Treball final de grau

GRAU DE MATEMÀTIQUES

Facultat de Matemàtiques i Informàtica
Universitat de Barcelona

Regressió a la mediocritat, ara i abans.

Autora: Júlia Aragó i Roca

Director: Dr. Josep Fortiana
Realitzat a: Departament de Matemàtiques i Informàtica

Barcelona, 18 de gener de 2019

Abstract

Regression towards mediocrity is a widely known statistical phenomenon, considered trivial when deeply understood. Regardless of its apparent simplicity, its interpretation seems unclear and continues to confuse people by producing fallacious reasoning.

This dissertation begins by showing a historical approach on the issue focusing on the origin of the regression theory and its most related topics: correlation and covariance.

Next, the mathematical bases which structure the regression method are shown from a geometrical point of view as well as its probabilistic equivalent. The Least Squares method, along with its historical motivation, is exposed as the most celebrated regression method.

After that, a discussion is presented on the subtlety of the regression method, which makes it be considered one of the most reproduced fallacy in the history of economic statistics and data visualization. Moreover, the most mistaken and surprising interpretations that have taken place since its appearance until nowadays are exposed. In order to achieve a good comprehension of the paradox, a simulation is run so the reader can relate the mentioned concepts with empirical data.

Finally, the relationship that binds the regression method with latter ones, such as Shrinkage or James-Stein estimator, is introduced. Such methods may be interpreted as an improvement of the regression method. As well as the method's rigorous explanation and proof, its Galtonian deduction is also referred.

Resum

La *Regressió a la Mediocritat* és un fenomen estadístic àmpliament conegut, considerat senzill quan és entès profundament. Malgrat la seva aparent simplicitat, la seva interpretació continua essent el principal problema en la justificació de resultats.

Aquest treball presenta una aproximació històrica sobre el problema de regressió entès com l'origen de la teoria de la regressió estadística, i els seus conceptes més propers: la correlació i la covariància.

Seguidament, es mostren les bases matemàtiques que estructuren el mètode de la regressió des d'un punt de vista geomètric i el seu equivalent més probabilístic. S'introdueix el mètode per excel·lència de l'anàlisi de regressió, el mètode dels mínims quadrats, així com la seva motivació històrica.

A continuació, es comenten i discuteixen les subtileses del mètode de regressió que fan que aquest es consideri una de les fal·làcies estadístiques més repetides al llarg de la història de l'estadística econòmica i la visualització de dades. Tanmateix, s'exposen les interpretacions errònies més sorprenents que s'han donat des de l'aparició del fenomen i fins l'actualitat. Es complementa la bona comprensió de la paradoxa amb una breu simulació que ajuda a visualitzar, amb dades palpables, els conceptes que s'han discutit.

Finalment, s'introdueix la relació que lliga el mètode amb altres mètodes posteriors com el Shrinkage o l'estimador James-Stein, el qual es pot interpretar com una versió millorada del mètode de regressió. De la mateixa manera que es mostra l'explicació rigorosa del mètode i la seva demostració, també es dona a conèixer la seva deducció "Galtoniana".

2010 Mathematics Subject Classification. 62A01, 62A86, 62J05, 62J07, 62J10, 62J86

Agraïments

A la meva família i amics, per haver-me acompanyat i recolzat en aquest procés que finalitza una etapa acadèmica i n'obre d'altres noves. Al meu tutor, per tota l'ajuda i acompanyament durant aquests mesos de recerca, sense el qual aquest treball no hagués estat possible. A tots ells, gràcies.

Índex

1 El naixement de l'Estadística moderna

1.1 Sir Francis Galton i la Llei de Regressió

Francis Galton va néixer el 16 de febrer de 1822 a Sparkbrook, Anglaterra. Era el menor de set germans d'una família acomodada. La seva etapa estudiantil va estar marcada per alternar els seus estudis en medicina i matemàtiques. Complint el desig del seu pare va iniciar els estudis de medicina, però el 1840 es trasllada al Trinity College de la Universitat de Cambridge per estudiar matemàtiques fins el 1844 quan, després de patir una crisi nerviosa, decideix tornar al camp de la medicina. Va ser aprenent a l'hospital General de Birmingham durant un temps per, finalment, continuar amb el seu treball matemàtic al King's College de Londres.

Arran de la mort del seu pare el 1844, Galton i els seus germans reben una generosa fortuna que els permet dedicar-se plenament a les seves aficions. El 1845 inicia un seguit de viatges arreu del món on realitza una sèrie d'observacions geogràfiques fins el 1850, moment en que retorna a Anglaterra i és guardonat amb la medalla d'or de la Royal Geographical Society. Més tard seria nomenat membre de la Royal Society.

Les grans aportacions de Galton a l'estadística moderna tenen una motivació lligada a l'estudi de la genètica, camp del qual va viure el moment més àlgid. Coetani de Gregor Mendel (1822-1884) i cosí Charles Darwin (1809-1882), Galton va quedar captivat per l'obra *Origin of the Species*, la qual el va portar a l'estudi de l'eugenèsia, que tracta la millora dels éssers humans a través de la reproducció selectiva.

És a finals del segle XIX quan Galton introdueix el concepte de *Regressió*. A l'article *Typical laws of heredity* 1877, que va exposar davant la Royal Institution de Londres en una de les seves trobades setmanals. En la seva intervenció, Galton va presentar les dades de l'estudi sobre l'heretabilitat de les característiques de les plantes de pèsol (sweet peas). Estudi que, més endavant, va concloure al llibre *Natural inheritance* del 1894.

Entre mig d'aquests dos estudis, i més proper al tema que és tracta en aquest treball, Galton va publicar el 1886 l'article *Regression towards mediocrity in hereditary stature*, on, mogut pel seu interès en l'estudi de l'heretabilitat i, en concret de les característiques humanes, físiques i mentals, va presentar el conjunt de dades que sostenen el que va anomenar *Llei de Regressió*.

Les dades consisteixen en el recull de les alçades de 928 fills adults i dels seus respectius progenitors. Galton tenia per objectiu estudiar la transferència genètica de l'alçada de pares a fills. Per fer-ho, va combinar l'alçada dels pares prenent la mitjana de les alçades dels dos progenitors i escalant les alçades de les dones multiplicades per 1.08 com a factor de correcció. D'aquesta manera va crear un total de 205 dades les quals va anomenar *midparents*. En elles hi va observar una relació lineal entre l'alçada dels pares i dels fills, podent prendre l'alçada del pare com a predictor de l'alçada del fill.

Més enllà d'aquesta relació lineal, Galton va presentar també l'anomenat *fenomen o efecte de regressió*. En l'estudi, es va descriure una relació de regressió entre les alçades dels pares i dels fills. Es ressaltava el fet que els fills de pares alts són alts però de mitjana, menys alts que els seus pares, i els fills de pares baixos són baixos però de mitjana menys baixos que els seus pares. Descobrint així una regressió a la mediocritat de les alçades dels fills respecte a les dels pares. Aquest fet dóna nom als anomenats avui en dia *model de regressió lineal* i *recta de regressió* (a la mitjana).

Aquesta regressió és deguda al propi mètode que tendeix a sobreestimar les respostes per als casos més extrems. És a dir, observacions extremes tendeixen a donar respostes més centrals (o properes a la mitjana global) i viceversa; observacions centrals tendeixen a donar respostes extremes. En cap cas significa que existeixi una convergència cap a una distribució central. En cas contrari, estaríem afirmant que l'alçada d'un determinat grup d'individus tendeix a estancar-se en una alçada constant h, fet que no s'ha observat mai.

Aquest fenomen es dóna quan les dades que s'estudien no guarden una relació estrictament funcional. En el cas de Galton l'alçada dels pares, variable X, i l'alçada dels fills, variable Y, pateixen, a més, factors aleatoris que fan que la variable Y, anomenada *resposta*, no es pugui donar com una funció determinista de la variable X, anomenada *predictiva* o *explicativa*. És a dir, l'alçada dels pares no és l'únic predictor de l'alçada dels fills, sinó que també hi intervenen altres factors aleatoris com l'alimentació o les malalties, per exemple.

La formalització matemàtica d'aquest fenomen tal i com es va desenvolupar històricament, es podria plantejar de la següent manera: Siguin $\{x_1, x_2, \ldots, x_n\}$ una primera observació d'una certa variable X (per exemple, l'alçada dels pares) i $\{y_1, y_2, \ldots, y_n\}$ una segona observació de la mateixa variable (per exemple, l'alçada dels fills), la regressió a la mitjana indica que per tot valor x_i, el valor esperat de y_i serà més proper al valor de $\overline{X}$ que al respectiu valor de x_i. D'altra manera,

$$\mathbb{E}(|(y_i - \overline{x})|) < \mathbb{E}(|(x_i - \overline{x})|),$$

per tot $i = 1, \ldots, n$.

1.2 L'el·lipse de Galton, correlació

Galton havia provat que l'efecte de regressió es dóna quan l'esdeveniment que estudiem és susceptible tant a influències deterministes com estocàstiques. Més concretament, és especialment fàcil d'observar en estudis longitudinals on s'observa una mateixa variable en dos moments diferents del temps. Però, el que és realment interessant a determinar és quin és aquest efecte de regressió i quina influència té sobre les forces deterministes.

Per a Galton l'interès i motivació del seu estudi es trobava en determinar quant influïa l'efecte de regressió sobre l'esperada de l'heretabilitat. Amb la finalitat d'exprimir tota la informació que les dades amagaven i mogut per un esperit molt

Cartesià, va provar de transformar les nombroses taules de dades que havia recollit sobre les alçades de les famílies en gràfics de dades més visuals. Havia creat el primer *scatterplot* o *diagrama de dispersió* i obert la porta a la ciència de visualització de dades.

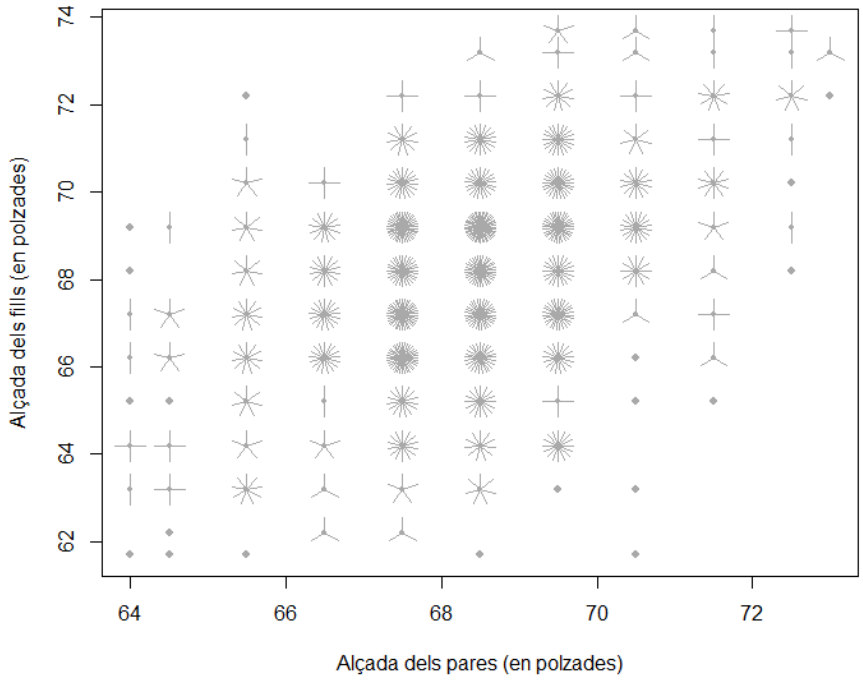


Figura 1: Sunflower Plot de les alçades dels fills contra les dels pares.

Com es pot veure, el núvol de punts format per les dades sembla agrupar-se en forma d'el·lipse. Galton va observar aquesta apreciació i va dibuixar corbes anomenades *isopletes* al llarg dels punts on la densitat era aproximadament constant, és a dir, *corbes de nivell*. Amb l'ajuda del matemàtic escocès J.H. Dickson (1849-1931), Galton va poder confirmar la seva sospita i dibuixar les el·lipses, com podem veure a la Figura 2. Aquestes corbes de nivell de Galton no van ser una invenció seva sinó que s'atribueixen a l'astrònom anglès Edmond Halley (1656-1742).

La particularitat que tenien les isopletes de Galton era que totes tenien forma d'el·lipses concèntriques, deixant intuir el que avui ja coneixem, una distribució Normal Bivariant de les dades. El que resulta realment interessant d'aquestes el·lipses és la relació geomètrica que guarden amb la correlació de les variables d'estudi. D'alguna manera Galton havia trobat la resposta a la gran pregunta de la Biologia avantguardista del segle XIX, la quantificació de l'heretabilitat, mitjançant el grau d'associació de dues variables. L'excentricitat de les el·lipses es va interpretar com la *correlació*. Com menys excèntriques eren les el·lipses (i en conseqüència més circular semblava el núvol de punts) menys associades estaven les variables i per tant, l'efecte de regressió tenia una forta influència en la predicció de les alçades

dels fills. En el cas contrari, com més excèntriques eren les el·lipses, més associades estaven les variables i més pes guanyava l'efecte de l'heretabilitat dels pares sobre els fills, deixant entreveure que no observaríem efecte de regressió davant d'un cas de correlació perfecte entre dues variables.

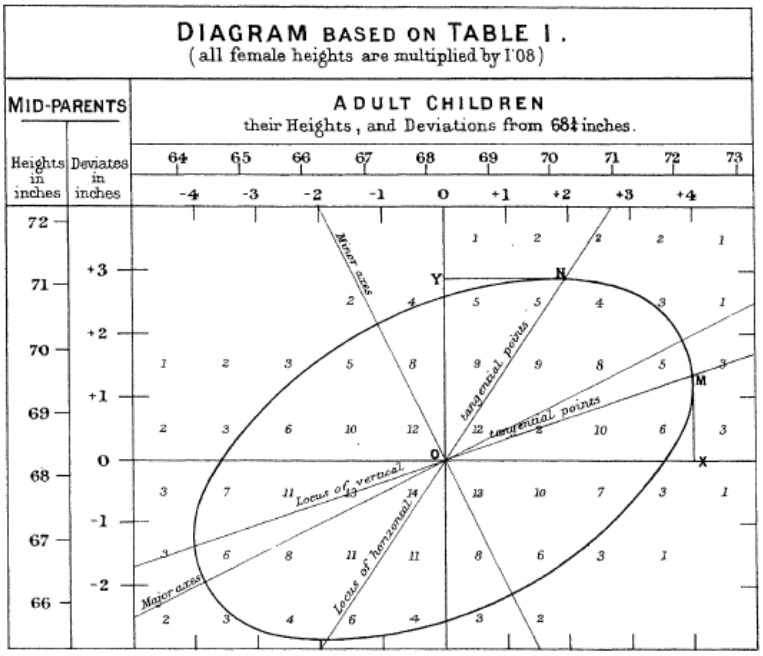


Figura 2: Primera Regressió a la mitjana feta per Galton amb les dades de les alçades dels pares i els fills (Galton, 1886 [9]).

La idea inicial de Galton va ser estesa per Udny Yule (1871-1951) i Karl Pearson (1857-1936), deixeble de Galton, qui va situar els descobriments del seu mestre a l'alçada dels de Charles Darwin.

El coeficient de correlació més usat avui en dia és, precisament, el de Pearson, que es coneix amb el nom de *coeficient de correlació lineal de Pearson.* Aquest coeficient és interpretat actualment com l'índex estadístic, r o ρ (de regressió), que mesura el grau d'associació lineal de dues variables quantitatives. Es defineix de la següent manera,

$$\rho = \frac{s_{xy}}{s_x s_y},$$

on

$$s_{xy} = \mathrm{cov}(X,Y) = \frac{\sum_{i=1}^{k}(x_i - \overline{x})(y_i - \overline{y})}{n}$$

i, s_x i s_y són les desviacions típiques de X i Y, respectivament, definides com l'arrel quadrada de la variància,

$$s_x = \sqrt{\frac{\sum_{i=1}^{k}(x_i - \overline{x})^2}{n}} \qquad \text{i} \qquad s_y = \sqrt{\frac{\sum_{i=1}^{k}(y_i - \overline{y})^2}{n}}.$$

Tot i que Galton només va contemplar el cas en que $0 < \rho < 1$, sabem que ρ està comprès entre $-1 \leq \rho \leq 1$. Si $\rho = |1|$ diem que existeix una correlació perfecte, que serà positiva o negativa segons el signe.

1.3 Galton i les dades dels pèsols

El 1875 Galton realitza el conegut experiment amb les plantes del pèsol d'olor (sweet peas). La raó de treballar amb aquestes plantes és, entre d'altres avantatges, la seva capacitat d'auto fertilitzar-se. D'aquesta manera les variacions genètiques de les plantes filles són únicament degudes a una planta mare, eliminant així el problema d'un segon progenitor. Galton va estudiar set grups de llavors diferents els quals va anomenar K, L, M, N, O, P i Q ordenats de major a menor pes, i va repartir-les entre diferents amics perquè li retornessin les plantes filles un cop haguessin germinat.

Posteriorment, va estudiar diverses característiques de 100 filles per cada llavor mare, aconseguint en total una mostra de 700 plantes filles. A continuació, va dibuixar la gràfica dels diàmetres de les llavors filles contra el de les llavors mare i va calcular les mitjanes de les 100 llavors filles provinents d'una mida concreta de llavors mare. A partir d'aquí, va poder observar que les mitjanes de les llavors filles de mateixa llavor mare descrivien aproximadament una recta amb pendent positiu menor que 1.

Aquest era el cas més senzill i especial, on la variació de les observacions de les llavors mares, variable x, és pràcticament la mateixa que la de les llavors filles, variable y, és a dir, $s_x = s_y$. Aleshores, en aquest cas tenim que el pendent de la recta de regressió, $\hat{\beta}$, coincideix amb el coeficient de correlació ja que, com veurem més endavant, es relacionen d'aquesta manera

$$\rho = \hat{\beta}\,\frac{s_x}{s_y} = \hat{\beta}\frac{\sqrt{\text{var}(x)}}{\sqrt{\text{var}(y)}},$$

per tant, si

$$\text{var}(x) = \text{var}(y) \implies \rho = \hat{\beta}.$$

Com es pot observar a la Figura 3, les mitjanes de cada columna tendeixen a alinear-se en una recta que podria considerar-se una aproximació grollera de la recta de regressió.

Per a les dades pèsol que va recollir Galton, tenim que

$$\begin{aligned} \text{var}(x) &= 4.005 \\ \text{var}(y) &= 3.925 \\ \hat{\beta} &= 0.342. \end{aligned}$$

L'aproximació que va donar Galton per al pendent de la recta de regressió va ser de 0.33, la qual és una molt bona aproximació.

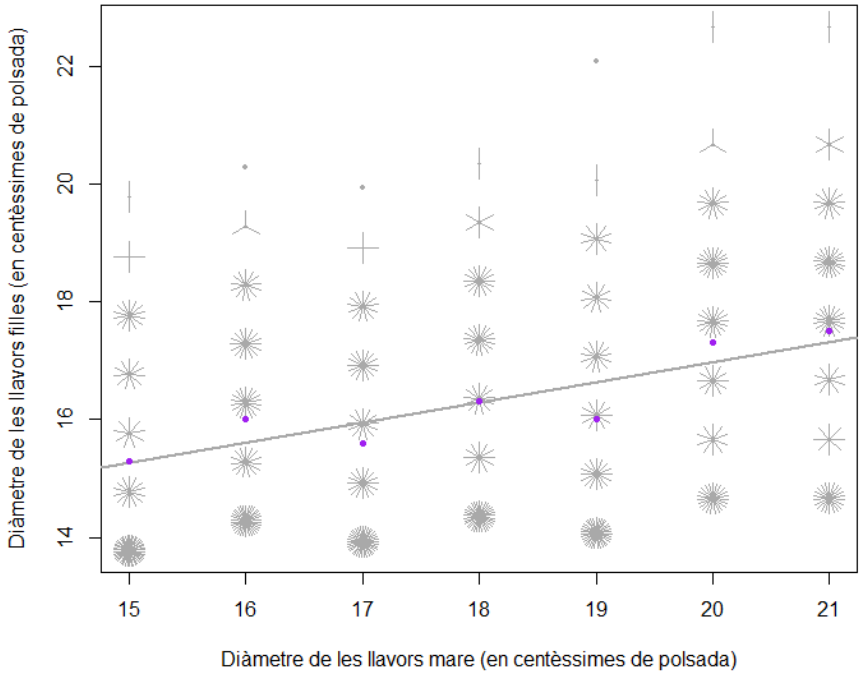


Figura 3: En gris, regressió de les 700 mesures dels diàmetres dels pèsols amb la seva recta de regressió. En lila, les mitjanes per cada columna corresponent a una llavor mare. (Veure annex 2.)

1.4 El Mètode dels Mínims Quadrats

El *mètode dels mínims quadrats* va ser el primer que es va usar en un anàlisi de regressió, i actualment es presenta als cursos d'Estadística com un mètode relacionat amb el problema de regressió o d'ajust. Aquest problema consisteix en trobar una corba que sota uns determinats criteris s'aproximi o s'"ajusti"de la millor manera possible als punts d'una distribució bivariant determinada.

No obstant, aquest mètode va sorgir motivat pel camp de l'astronomia com a tècnica geodèsica. Són diversos els antecedents allunyats de l'estadística que van generar el mètode: el problema de la figura de la Terra, que va desencadenar en la necessitat de mesurar la longitud de l'arc meridional terrestre, que alhora va introduir el sistema mètric decimal.

Diversos autors van publicar els seus respectius descobriments d'aquests problemes durant el primer terç del segle XIX, fet que va donar lloc a agres disputes sobre l'autoria del mètode.

1.4.1 La primera publicació, Legendre

El mètode va ser publicat per primera vegada el 1805 per Adrien-Marie Legendre (1752-1833) en un apèndix del llibre sobre les òrbites dels cometes *Nouvelles méthodes pour la détermination des orbites des cométes*.

L' 1 de gener del 1801 l'astrònom italià Giusseppe Piazzi (1746-1826) de l'observatori de Palerm descobreix l'asteroide Ceres, el més gran del sistema solar. Degut a la seva mala situació respecte al Sol, només se'n van poder fer observacions durant 40 dies abans de perdre'l. Aleshores, Piazzi va publicar les dades de les poques observacions que tenia amb l'esperança que altres científics poguessin determinar la seva trajectòria. Va ser llavors quan Carl Friedrich Gauss (1777-1855) va publicar, sense donar masses explicacions, un mètode que predeia l'òrbita de l'asteroide. Immediatament després es va poder trobar Ceres quan aquest apareixia per l'altra banda del Sol, exactament on Gauss havia predit.

Tot i suposar-se que Gauss era ja coneixedor del mètode dels mínims quadrats, no va publicar-lo fins el 1809 a *Theoria motus corporum coelestium* on, a més, Gauss en reclama l'autoria i assegura estar utilitzant-lo des de 1795, obrint així la disputa amb Legendre (Stigler, 1977 [29]).

1.4.2 Gauss en reclama l'autoria

Diverses són les causes que fan pensar que, si bé Gauss no va ser el primer en desenvolupar el mètode, almenys hi va arribar paral·lelament al treball de Legendre. Principalment tenim 4 proves existents:

i. La reclama de Gauss a la seva publicació del 1809 afirmant que va utilitzar el mètode per predir l'òrbita de Ceres.

ii. Una críptica entrada al seu diari matemàtic del 17 de juny del 1798: *"Calculus probabilitatis contra La Place defensus"*. ("El càlcul de probabilitat defensat en contra de Laplace")

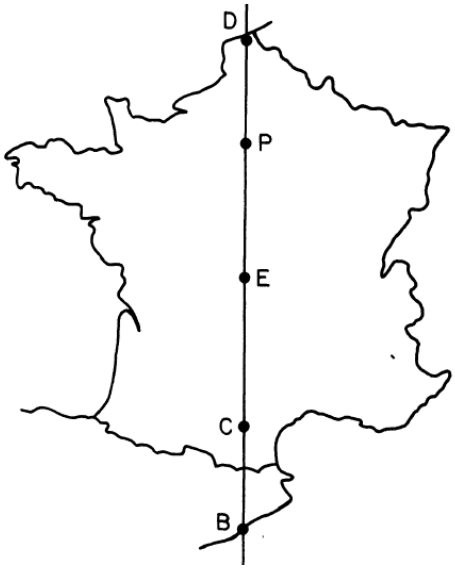
iii. L'afirmació de Gauss d'haver comunicat el mètode a altres astrònoms com Oblers (1758-1840), Lindenau (1780-1854) i von Zach (1754-1832), als quals va sol·licitar el seu testimoni.

iv. Una carta de Gauss publicada el 1799 en *Allgemeine Geographische Ephemeriden* on es refereix al "meine Method" ("meu mètode") en referència al problema de mesura de l'arc meridional.

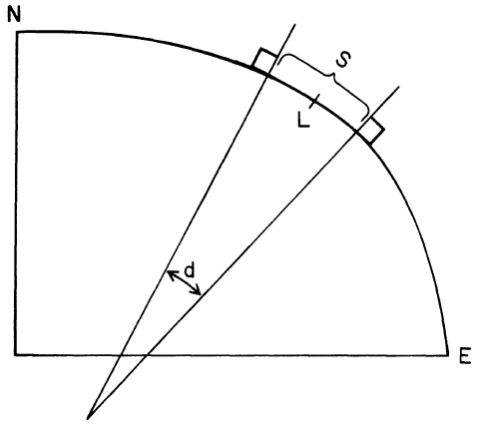
Les tres primeres afirmacions constitueixen una defensa força feble al seu favor però, és la quarta afirmació el que realment fa pensar que Gauss ja era coneixedor del mètode des d'abans de la publicació de Legendre.

Aquest problema és força conegut ja que va ser el que va portar a la concreció del primer metre i el sistema mètric decimal. Durant la segona meitat del segle XVIII,

l'acadèmia de les ciències de Paris perseguia la idea de determinar un sistema mètric comú basat en una nova mesura, el metre, equivalent a una deu milionèsima part del quadrant meridional. Per fer-ho però, s'havia de mesurar el meridià. L'elegit va ser el que passa per Paris però, davant la impossibilitat de mesurar-lo sencer es va decidir mesurar l'arc de meridià comprès entre Dunkerque i Barcelona en quatre parts.



Arc meridià francès, per Dunkerque (D), el panteó de Paris (P), Evaus (E), Carcassona (C) i Barcelona (B).



Quadrant meridià, des de l'Equador (E) fins al Pol nord (N).

Figura 4: Stigler, 1981 [30].

La complicació del problema requeia en el fet que les relacions entre la longitud d'arc, l'excentricitat i el quadrant meridional no són lineals. Existeixen moltes maneres de convertir el problema en un problema lineal de mínims quadrats, com les proposades per Boscovich el 1755, Laplace el 1780 o Legendre el 1805 entre d'altres,

$$a = z + y \sin^2 L$$

aquesta és una bona aproximació per a arcs petits considerant la Terra el·lipsoïdal, on $a = S/d$ és la longitud en mòdul per graus, z és la longitud d'un grau a l'equador i y és la diferència de graus al pol respecte a l'equador.

Cap d'aquests mètodes però, no dóna un resultat proper al que va donar Gauss. Per tant, podem pensar que, o bé Gauss va cometre errors de càlcul o bé no va utilitzar cap forma lineal de mínims quadrats. La primera opció és fàcilment descartable ja que Gauss és conegut com un gran calculista que poc probablement hauria comès un error tan gran en tants pocs passos. Així doncs, Gauss no hauria usat mínims quadrats purs. Com explica Stigler (1981) [30], el més probable és que Gauss utilitzés una aproximació de segon ordre per mínims quadrats per obtenir els seus resultats, molts propers als de Boscovich. En conseqüència, podem creure amb força certesa que Gauss va descobrir el mètode entre 1794 i 1799.

Si bé és possible que Gauss arribés abans al descobriment del mètode, i fins i tot anés més enllà que Legendre relacionant-lo als coneixements probabilístics amb mencions a la distribució Normal com a distribució dels errors, qui realment va saber transmetre i comunicar la seva importància i situar-lo en primer pla del panorama científic va ser Legendre.

1.4.3 Altres publicacions

Tot i que en parlar de la retrospectiva històrica del mètode dels mínims quadrats la disputa entre Gauss i Legendre s'emporta tot el protagonisme, les publicacions del mètode són diverses.

Cal destacar la publicació de Boscovich (1711-1787) que va proposar el seu propi mètode el 1755 el qual Laplace (1749-1827) (qui també va proposar-ne un el 1780) va batejar com *"mètode de situació"* per diferenciar-lo del mètode de mínims quadrats. A diferència del mètode de Legendre de 1805, que minimitza la suma dels quadrats dels errors, el mètode proposat per Boscovich minimitza la suma del valor absolut dels errors.

2 Sobre la Regressió lineal

2.1 Problema de Regressió Lineal

Deixant enrere la contextualització històrica, és convenient fer un petit recordatori dels conceptes que s'han tractat des d'un enfocament més teòric:

Donades dues variables aleatòries X i Y, com podrien ser en el treball de Galton l'alçada dels pares i els fills, respectivament, podem plantejar-nos el *problema de regressió* associat. Aquest consisteix en determinar relacions entre X i Y que ens permetin predir el valor d'una d'elles en funció de l'altra. Aquestes relacions poden ser diverses, en aquest punt ens centrarem en el cas lineal, així com passa al treball de Galton. D'aquesta manera, podem escriure aquesta relació de la forma

$$y = \alpha + \beta x + \varepsilon,$$

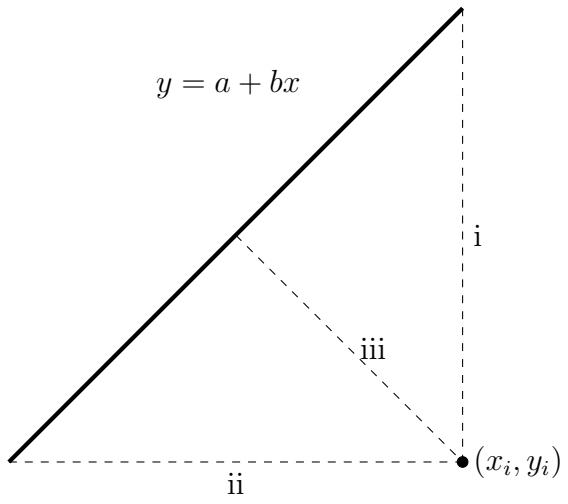
on $\varepsilon = \varepsilon(x, y)$ indica la pertorbació, la part estocàstica o no determinista.

L'objectiu del problema de regressió lineal és trobar quina recta $y = \hat{\alpha} + \hat{\beta}x$ s'ajusta millor a les dades. Donat un conjunt de dades, existeixen moltes maneres per determinar quina és aquesta recta. Per tant, el que hem d'especificar és una funció distància $F(\hat{\alpha}, \hat{\beta})$ entre la recta i el conjunt de dades, i trobar els $\hat{\alpha}$ i $\hat{\beta}$ que minimitzen aquesta funció.

Donat un punt (x_i, y_i) qualsevol, podem definir la distància d'aquest punt a la recta, d_i. Per a un vector aleatori (X, Y) de dimensió n tenim un conjunt de n distàncies $\{d_1, ..., d_n\}$. Prenem F com a representant de les n distàncies. Aleshores, hem de determinar:

- Quina distància d_i prenem,
- Com obtenim el representant F del conjunt $\{d_1, ...d_n\}$.

Hi ha diverses possibles distàncies d_i que podríem triar, per exemple:



i. Distància vertical, ens permet obtenir la recta de regressió de Y sobre X.

ii. Distància horitzontal, ens permet obtenir la recta de regressió de X sobre Y.

iii. Distància ortogonal, ens permet obtenir el primer eix principal.

Pel que fa a l'elecció del representant F, també tenim diferents opcions. Podríem triar la mitjana de les distàncies individuals

$$\overline{d} = \frac{1}{n}\sum_{i=1}^{n} d_i.$$

D'aquesta manera però, podria passar que $\exists\ i, j,\ i \neq j$ tals que $d_i = -d_j$, perdent així informació sobre la dispersió dels punts. Per intentar solucionar aquest problema podríem definir F com

$$|\overline{d}| = \sum_{i=1}^{n} |d_i|$$

o bé escollir la mediana del conjunt de distàncies $\{d_i, 1 \leq i \leq n\}$. Tot i així, aquestes dues opcions tampoc serien les més adequades, ja que no són funcions derivables i, per tant, no són bones candidates a l'hora de minimitzar.

Una funció, entre d'altres, que compleix aquests requeriments és la mitjana de les distàncies individuals al quadrat

$$F(\alpha, \beta) = \overline{d^2} = \frac{1}{n}\sum_{i=1}^{n} d_i^2,$$

amb aquesta definició, F és una funció contínua i derivable i, per tant, fàcil de minimitzar.

2.2 Regressió per Mínims Quadrats

Amb l'objectiu d'estudiar la recta de regressió de Y/X triarem com a distància individual d_i la distància vertical de cada punt a la recta, anomenada *error* o *residu*, és a dir,

$$d_i = y_i - \hat{y}_i = y_i - (\alpha + \beta x_i).$$

El mètode de mínims quadrats suggereix que els paràmetres α i β s'haurien de determinar triant els valors que minimitzen la suma dels quadrats dels errors, donada per

$$F(\alpha, \beta) = \overline{d^2} = \frac{1}{n}\sum_{i=1}^{n} d_i^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - (\alpha + \beta x_i))^2.$$

Per tal de realitzar els càlculs, recordem que les variables X i Y tenen per mitjanes

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad \text{i} \qquad \overline{xy} = \sum_{i=1}^{n} x_i y_i,$$

les mitjanes dels seus quadrats són

$$\overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 \qquad \text{i} \qquad \overline{y^2} = \frac{1}{n}\sum_{i=1}^{n} y_i^2,$$

i les seves variàncies i covariàncies són,

$$s_x = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \overline{x^2} - \overline{x}^2, \qquad s_y = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \overline{y^2} - \overline{y}^2$$

i

$$s_{xy} = \frac{1}{n}\sum_{i=0}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \overline{xy} - \overline{x}\,\overline{y}.$$

Aleshores tenim que,

$$\begin{aligned} F(\alpha,\beta) &= \frac{1}{n}\sum_{i=1}^{n}(y_i^2 + \alpha^2 + \beta^2 x_i^2 - 2\alpha y_i - 2\beta x_i y_i + 2\alpha\beta x_i) = \\ &= \overline{y^2} + \alpha^2 + \beta^2\overline{x^2} - 2\alpha\overline{y} - 2\beta\overline{xy} + 2\alpha\beta\overline{x} = \\ &= (\overline{y^2} - \overline{y}^2) + \overline{y}^2 + \alpha^2 + (\beta^2\overline{x^2} - \beta^2\overline{x}^2) + \beta^2\overline{x}^2 - 2\alpha\overline{y} - (2\beta\overline{xy} + 2\beta\overline{xy}) - 2\beta\overline{xy} + 2\alpha\beta\overline{x} = \\ &= s_y^2 + \beta^2 s_x^2 - 2\beta s_{xy} + (\overline{y} - \alpha - \beta\overline{x})^2. \end{aligned}$$

Hem de calcular els α i β que minimitzen aquesta funció. En primer lloc, l'últim sumand no serà mai negatiu per tant, per qualsevol β podem fer que sigui zero prenent

$$\alpha = \hat{\alpha}(\beta) = \overline{y} - \beta\overline{x}.$$

Fixant $\alpha = \hat{\alpha}(\beta)$, podem trobar quin β fa mínim $F(\hat{\alpha}(\beta),\beta) = s_y^2 + \beta^2 s_x^2 - 2\beta s_{xy}$. Derivant respecte de β, i igualant a zero, obtenim

$$\hat{\beta} = \frac{s_{xy}}{s_x^2}.$$

Finalment, substituint obtenim l'equació de la *recta de regressió de Y sobre X*

$$y = \overline{y} - \frac{s_{xy}}{s_x^2}\,\overline{x} + \frac{s_{xy}}{s_x^2}\,x.$$

És fàcil comprovar que les solucions $\hat{\alpha}$ i $\hat{\beta}$ trobades són mínims de F ja que la segona derivada respecte $\hat{\beta}$ és $2s_x \geq 0$ per tot x.

2.3 Regressió Lineal Múltiple

L'extensió natural del problema de regressió lineal és la *Regressió lineal múltiple.* A diferència de la regressió lineal, que estudia la relació entre una variable dependent Y i una variable explicativa X, a la pràctica és habitual trobar-nos que la variable Y depengui de més d'una variable explicativa.

D'aquesta manera, l'ampliació del problema per a $n > 2$ punts $(x_i, y_i) \in \mathbb{R}^{p+1}$, on $x_i \in \mathbb{R}^p, p \geq 1$ és un vector fila de dimensió $(1 \times p)$ que escriurem com $x_i = \begin{pmatrix} x_{i1}, & x_{i2}, & \dots, & x_{ip} \end{pmatrix}$ i $y_i \in \mathbb{R}, 1 \leq i \leq n$ vector columna $(n \times 1)$

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},$$

consisteix a estimar els paràmetres $\alpha \in \mathbb{R}$ i $\beta \in \mathbb{R}^p$ que escriurem com un vector columna $(n \times 1)$

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix},$$

tals que minimitzin la mateixa funció de regressió

$$F(\alpha, \beta) = \overline{d^2} = \frac{1}{n} \sum_{i=0}^{n} d_i^2,$$

amb $d_i = y_i - \hat{y}_i = y_i - \alpha - x_i \cdot \beta,\ 1 \leq i \leq n.$

El problema matricial associat és el següent:

$$F(\alpha, \beta) = \frac{1}{n} \sum_{i=n}^{n} \left[\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} - \begin{pmatrix} \alpha \\ \alpha \\ \vdots \\ \alpha \end{pmatrix} - \begin{pmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1p} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & x_{n3} & \dots & x_{np} \end{pmatrix} \cdot \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix} \right]^2.$$

Els càlculs venen donats de la mateixa manera que al cas lineal amb alguns canvis de notació:

$$F(\alpha, \beta) = s_y^2 + \beta' \cdot S_x \cdot \beta - 2S_{xy} \cdot \beta + (\overline{y} - \alpha - \overline{x} \cdot \beta)^2,$$

on s_y^2 és la variància de y, S_x és la matriu de variàncies i covariàncies de la variable X

$$S_x = \frac{1}{n} \sum_{i=0}^{n} (x_i - \overline{x})'(x_i - \overline{x}) = Q_x - \overline{x}' \cdot \overline{x},$$

tal que

$$Q_x = \frac{1}{n}\sum_{i=0}^{n} x_i' \cdot x_i,$$

és una matriu $(p \times p)$ i $\overline{x}$ és el vector fila $(1 \times p)$ de mitjanes, i S_{xy} és el vector fila $(1 \times p)$ que té per components les covariàncies de les components de x amb y.

Finalment, podem trobar la solució del problema seguint el mateix procediment que al cas lineal; primerament posant a zero l'últim sumand prenent $\hat{\alpha} = \overline{y} - \overline{x} \cdot \beta$, i minimitzant a continuació la resta de sumands que només depenen de β. Derivant i igualant a zero trobem,

$$S_x \cdot \beta = S_{xy}'.$$

Aquesta equació té solució única quan S_x no és singular:

$$\hat{\beta} = S_x^{-1} \cdot S_{xy}'.$$

En conseqüència, substituint aquesta solució, l'equació de l'hiperplà de regressió obtingut és

$$Y = \overline{y} - \overline{x} \cdot S_x^{-1} \cdot S_{xy}' + X \cdot S_x^{-1} \cdot S_{xy}'.$$

2.4 Versió Probabilística

Fins ara hem vist quin és el Problema de regressió i una de les seves solucions més freqüents; la recta de regressió (quan el problema es planteja a $\mathbb{R}^2$) o l'hiperplà de regressió (a $\mathbb{R}^{p+1}$) de Y sobre X, obtinguts amb el mètode dels mínims quadrats. Com ja hem dit anteriorment, aquesta és la solució que trobem quan les variables que estudiem admeten una relació lineal, però això no és sempre així.

Donat un vector aleatori (X, Y) amb Probabilitat Bivariant $H(x, y)$ tal que X és la marginal de la probabilitat conjunta, tractar d'identificar les relacions que hi pugui haver entre les variables és equivalent a descriure una funció de regressió $Y = G(X)$ que ens descrigui la llei de Y, on $G(x) = \mathbb{E}(Y|X = x)$. En construir aquesta relació $G(X)$ estem considerant una corba de solucions al pla $y = G(x)$. D'aquesta manera, un punt qualsevol de la distribució (x, y) es transforma en $(x, G(x))$. La bondat de l'ajust d'aquesta corba recau en determinar $Y = G(X)$ tal que

$$\mathbb{E}((Y - G(X))^2) \quad \text{sigui mínima.}$$

Així, la corba $y = \mathbb{E}(Y|X = x)$ corresponent serà la que millor s'ajusti a les dades en quant a minimitzar els errors. Aquesta corba s'anomena *Corba de regressió* (a la mitjana) de Y sobre X per mínims quadrats.

2.4.1 Cas lineal

Podem veure que si aquesta corba és una recta, és a dir, si expressem la funció de regressió com una funció lineal $G(X) = a + bX$, aleshores aquesta corba coincideix exactament amb la recta de regressió.

Sigui (X, Y) un vector aleatori amb distribució conjunta que té moments fins a segons ordre. Volem resoldre el problema de regressió esmentat al punt anterior, trobar una variable aleatòria $\hat{Y} = \alpha + \beta X$ tal que $\mathbb{E}(|Y - \hat{Y}|^2)$ sigui mínima.

$$\begin{aligned}
\mathbb{E}(|Y - \hat{Y}|^2) &= \mathbb{E}((Y - \alpha - \beta X)^2) \\
&= \mathbb{E}(Y^2 + \alpha^2 + \beta^2 X^2 - 2\alpha Y - 2\beta XY + 2\alpha\beta X) \\
&= \mathbb{E}(Y^2) + \alpha^2 + \beta^2 \mathbb{E}(X^2) - 2\alpha\mathbb{E}(Y) - 2\beta\mathbb{E}(XY) + 2\alpha\beta\mathbb{E}(X) \\
&\overset{(*)}{=} \operatorname{var}(Y) + \mathbb{E}(Y)^2 + \alpha^2 + \beta^2 \operatorname{var}(X) + \beta^2\mathbb{E}(X)^2 - 2\alpha\mathbb{E}(Y) \\
&- 2\beta \operatorname{cov}(X, Y) - 2\beta\mathbb{E}(X)\mathbb{E}(Y) + 2\alpha\beta\mathbb{E}(X) \\
&= \operatorname{var}(Y) + \beta^2 \operatorname{var}(X) - 2\beta \operatorname{cov}(X, Y) + (\mathbb{E}(Y) - \alpha - \beta\mathbb{E}(X))^2.
\end{aligned}$$

En la quarta $(*)$ igualtat hem usat que

$$\begin{aligned}
\operatorname{var}(X) &= \mathbb{E}(X^2) - \mathbb{E}(X)^2 \\
\operatorname{var}(Y) &= \mathbb{E}(Y^2) - \mathbb{E}(Y)^2 \\
\operatorname{cov}(X, Y) &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).
\end{aligned}$$

Aleshores hem obtingut

$$\mathbb{E}(|Y - \hat{Y}|^2) = \operatorname{var}(Y) + \beta^2 \operatorname{var}(X) - 2\beta \operatorname{cov}(X, Y) + (\mathbb{E}(Y) - \alpha - \beta\mathbb{E}(X))^2.$$

L'últim sumand és estrictament positiu si és diferent de zero, per tant podem prendre $\hat{\alpha} = \mathbb{E}(Y) - \beta\, \mathbb{E}(X)$ i minimitzar

$$\mathbb{E}(|Y - \hat{Y}|^2) = \operatorname{var}(Y) + \beta^2 \operatorname{var}(X) - 2\beta \operatorname{cov}(X, Y).$$

Derivant respecte β i igualant a zero obtenim

$$\hat{\beta} = \frac{\operatorname{cov}(X, Y)}{\operatorname{var}(X)},$$

i substituint

$$\hat{\alpha} = \mathbb{E}(Y) - \frac{\operatorname{cov}(X, Y)}{\operatorname{var}(X)}\mathbb{E}(X).$$

Podem comprovar que les solucions $\hat{\alpha}$ i $\hat{\beta}$ són mínims ja que la segona derivada respecte $\hat{\beta}$ és $2\operatorname{var}(X) \geq 0$ per tot x.

D'igual manera que al punt anterior, obtenim novament l'equació de la recta de regressió de Y/X,

$$y = \mathbb{E}(Y) - \frac{\operatorname{cov}(X, Y)}{\operatorname{var}(X)}\,\mathbb{E}(X) + \frac{\operatorname{cov}(X, Y)}{\operatorname{var}(X)}\,x.$$

2.5 Distribució Normal Bivariant

És interessant veure com molts dels resultats comentats fins al moment es poden obtenir estudiant la forma funcional de la distribució Normal Bivariant. Ja que per a un vector aleatori (X, Y) amb distribució Normal Bivariant la solució de la corba de regressió és la recta de regressió, és a dir, la corba de regressió a la mitjana és una recta.

Siguin X i Y dues variables aleatòries amb distribució conjunta Normal Bivariant, amb mitjanes

$$\mathbb{E}(X) = \mu_x \quad i \quad \mathbb{E}(Y) = \mu_y,$$

variàncies

$$\text{var}(X) = \sigma_x \quad i \quad \text{var}(Y) = \sigma_y$$

i covariància

$$\text{cov}(X, Y) = \sigma_{xy} = \rho\sigma_x\sigma_y.$$

Notem que hem expressat la covariància en termes del coeficient de correlació lineal

$$\rho = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)}\sqrt{\text{var}(Y)}}.$$

Aleshores, la distribució conjunta Bivariant ve donada per

$$f_{(X,Y)}(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp Q(x, y)$$

on

$$Q(x, y) = \frac{-1}{2(1-\rho^2)} \left\{ \left(\frac{x-\mu_x}{\sigma_x} \right)^2 + \left(\frac{y-\mu_y}{\sigma_y} \right)^2 - 2\rho \left(\frac{x-\mu_x}{\sigma_x} \right) \left(\frac{y-\mu_y}{\sigma_y} \right) \right\}.$$

Per definició podem determinar la funció de densitat condicionada com

$$f_{Y|X=x}(y) = \frac{f(x, y)}{f_x(x)}.$$

Per tal de facilitar els càlculs podem expressar l'exponent $Q(x, y)$ de la densitat conjunta com

$$Q(x,y) = \frac{-1}{2(1-\rho^2)} \left\{ \left(\frac{x-\mu_x}{\sigma_x} \right)^2 + \right.$$
$$\left. + \left[\left(\frac{y-\mu_y}{\sigma_y} \right)^2 - 2\rho \left(\frac{x-\mu_x}{\sigma_x} \right) \left(\frac{y-\mu_y}{\sigma_y} \right) + \rho^2 \left(\frac{x-\mu_x}{\sigma_x} \right)^2 \right] - \rho^2 \left(\frac{x-\mu_x}{\sigma_x} \right)^2 \right\} =$$
$$= \frac{-1}{2(1-\rho^2)} \left\{ \left(\frac{x-\mu_x}{\sigma_x} \right)^2 + \left[\left(\frac{y-\mu_y}{\sigma_y} \right) - \rho \left(\frac{x-\mu_x}{\sigma_x} \right) \right]^2 - \rho^2 \left(\frac{x-\mu_x}{\sigma_x} \right)^2 \right\} =$$
$$= \frac{-1}{2(1-\rho^2)} \left\{ (1-\rho^2) \left(\frac{x-\mu_x}{\sigma_x} \right)^2 + \left[\left(\frac{y-\mu_y}{\sigma_y} \right) - \rho \left(\frac{x-\mu_x}{\sigma_x} \right) \right]^2 \right\} =$$
$$= -\frac{1}{2} \left(\frac{x-\mu_x}{\sigma_x} \right)^2 - \left\{ \frac{1}{2\sqrt{(1-\rho^2)}} \left[\left(\frac{y-\mu_y}{\sigma_y} \right) - \rho \left(\frac{x-\mu_x}{\sigma_x} \right) \right] \right\}^2 .$$

Donat que $f_x(x)$ és la marginal de $X \sim N(\mu_x, \sigma_x^2)$ tenim que

$$f_X(x) = \frac{1}{\sigma_x \sqrt{2\pi}} \exp\left(-\frac{(x-\mu_x)^2}{2\sigma_x^2} \right),$$

i a partir de la definició de la densitat condicionada, obtenim que

$$f_{Y|X=x}(y) = \frac{1}{\sigma_y \sqrt{2\pi(1-\rho^2)}} \exp\left\{ -\frac{1}{2\sqrt{(1-\rho^2)}} \left[\left(\frac{y-\mu_y}{\sigma_y} \right) - \rho \left(\frac{x-\mu_x}{\sigma_x} \right) \right] \right\}^2 .$$

Manipulant algebraicament l'exponent obtingut d'aquesta densitat condicionada, podem expressar-lo de la següent manera,

$$-\frac{1}{2(1-\rho^2)} \left(\frac{y - [\mu_y - \rho\frac{\sigma_y}{\sigma_x}(x-\mu_x)]}{\sigma_y} \right)^2 = -\frac{1}{2(1-\rho^2)} \left(\frac{(y-\mu_{y|x})^2}{\sigma_x^2} \right).$$

I, per tant, finalment obtenim

$$f_{Y|X=x}(y) = \frac{1}{\sigma_y \sqrt{2\pi(1-\rho^2)}} \exp\left(-\frac{(y-\mu_{y|x})^2}{2\sigma_y^2(1-\rho^2)} \right).$$

Com podem observar, la densitat de y condicionada a $X = x$ es comporta com una normal $N(\mu_{y|x}, \sigma_y^2)$. Per tant, l'esperança d'aquesta densitat és,

$$\mathbb{E}(f(y|x)) = \mu_{y|x} = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(x-\mu_x) = \mu_y + \frac{\sigma_{xy}}{\sigma_x^2}(x-\mu_x)$$

que és exactament la recta de regressió que hem definit al punt 2.2 amb una notació diferent.

3 La Fal·làcia de la Regressió

3.1 Horace Secrist i el triomf de la fal·làcia

Horace Secrist (1881-1943) va ser un estadístic i economista americà professor de la universitat de Northwestern i director del *Bureau of Economic Research* ("Oficina de recerca econòmica") de la mateixa universitat. El 1918 va passar a formar part de la *American Statistical Association*. Amb nombroses publicacions i ocupant importants càrrecs pel govern federal, Secrist era un reconegut investigador dins de la comunitat científica.

El 1933, a l'edat de 51 anys, Horace Secrist publica la culminació de 10 anys de recerca i investigació duta a terme per ell i el seu equip de més de 45 investigadors. El llibre *The Triumph of Mediocrity in Business* apareix en un moment de vital necessitat en el que, després de la Gran Depressió, es trobava en una posició idònia per diagnosticar i subministrar la cura per a una malferida economia nacional i internacional. El llibre era imponent i enormement detallat. Unes 468 pàgines que constaven de 140 taules i 103 gràfiques, totes i cadascuna d'elles acuradament documentades i explicades. Hom es podria preguntar que, si la majestuositat d'aquest llibre és, en nombres, equiparable a la de *Origin of Species* de Darwin, per què no ha assolit la mateixa fama?

La resposta és ben clara, com va dir Hotelling (ho veurem detalladament més endavant): "El que és interessant és incorrecte, i el que és correcte és trivial".

El llibre Secrist anuncia el que hauria de ser el descobriment més enlluernador de la teoria econòmica moderna. En les seves pròpies paraules:

> Mediocrity tends to prevail in the conduct of competitive business [...]
> Such is the price which industrial freedom brings.

Secrist havia descobert que la mediocritat era l'estat final al qual tendien a convergir els negocis i aportava nombroses proves que donaven suport a aquesta llei, com ara dades sobre els beneficis de 49 grans magatzems de la dècada del 1920 al 1930. Va seguir durant 10 anys les fortunes de les botigues per veure com responien respecte a l'estat econòmic inicial. Va dividir les 49 observacions en quatre quartils en funció de la riquesa inicial, i va anar seguint l'evolució de les mitjanes dels grups durant els següents 10 anys. Aquest fet el va portar a observar una clara tendència cap a la mitjana global de les dades, cap a la mediocritat, com es pot veure reflectit a la Figura 5.

Va observar i estudiar les dades de totes les maneres possibles i va arribar a la conclusió que no era un fenomen temporal marcat per l'any d'inici. Prengués la data inicial on fos que la prengués, la convergència a la mediocritat era sempre present. Davant d'aquestes evidències, Secrist va citar Galton per expressar les seves conclusions en la mateixa terminologia que va aplicar al procés de l'herència,

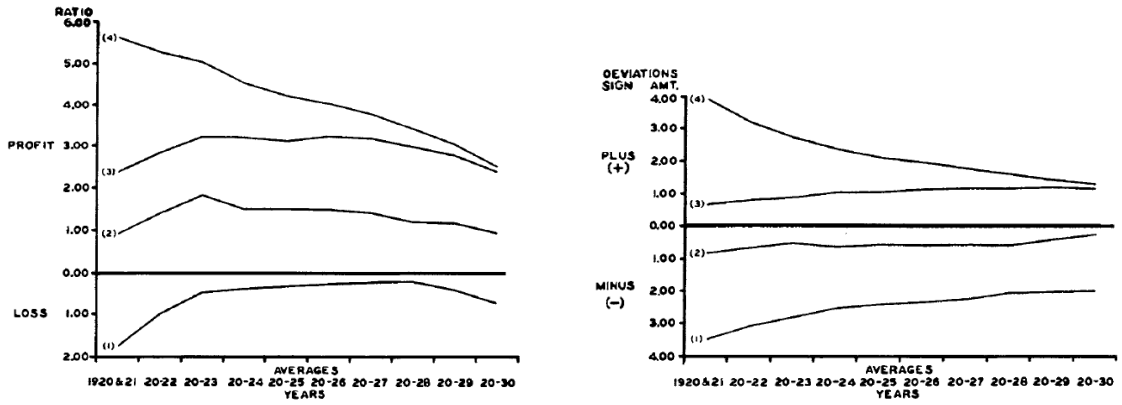


Figura 5: Dues de les 104 gràfiques presentades per Secrist, on es mostra la tendència de les mitjanes al llarg dels anys dels 49 grans magatzems agrupats en funció del seu valor al 1920.

> Both expenses and profits approach the mean, or to use Sir Francis Galton's expression, 'regress to type'.

Aquest remarcable descobriment hagués produït efectes immediats sobre l'economia internacional. Secrist va aconseguir captivar l'atenció del panorama econòmic de l'època.

Davant d'aquesta descoberta és natural preguntar-se si és possible que al darrere s'hi amagui un "accident estadístic"degut a la temporalitat de les dades, o bé que es tracti d'un "artefacte"o "defecte estadístic"derivat de la tipologia de les dades. El mateix Secrist va preocupar-se d'estudiar aquestes qüestions.

Per descartar que aquest fenomen pogués estar lligat a la naturalesa de sèrie temporal de les dades, que hagués pogut donar lloc a un accident estadístic, Secrist va estudiar un total de 73 sèries de dades de tot tipus d'empreses: botigues de queviures, ferreteries, empreses ferroviàries o bancs. Va observar-los des de qualsevol perspectiva temporal possible, obtenint sempre els mateixos resultats. La regressió a la mediocritat era una norma universal per als negocis americans.

Faltava descartar, però, que els resultats no es deguessin a la tipologia de les dades. Observaria el mateix si estudiava dades que no fossin econòmiques? Amb aquesta finalitat va analitzar una sèrie de 10 anys de dades de les temperatures mitjanes de juliol de 191 ciutats americanes, agrupades de la mateixa manera descrita pels grans magatzems. Aquesta vegada, lluny de trobar regressió, va trobar estabilitat.

Tota aquesta informació va portar a Secrist a creure que el que motivava la regressió de les dades econòmiques era la competitivitat dels mercats econòmics, concloent que la regressió sorgia allà on prevalien les forces de competitivitat motivades pel control humà.

Les primeres crítiques del llibre van ser favorables. En destaca la de la *Royal Statistical Society* ("Current notes" Vol. 96, No. 4, pp. 721-722) que va publicar una sinopsi dels descobriments amb grans felicitacions i reconeixement cap a Secrist

i el seu equip. També va rebre bones ressenyes de *The American Economic Review* (Elder, 1934 [6]), del *Journal of Political Economy* (King, 1934 [18]) i de *Annals of the American Academy of Political and Social Science* (Riegel, 1933 [23]).

Però l'estat de glòria que experimentava Secrist amb la que hauria de convertir-se en la seva obra mestra no va trigar massa en esvair-se. Estava a punt de conèixer la ressenya que tiraria per terra 10 anys d'intens treball dedicat, que posaria de manifest que la aparent convergència, elevada al nivell de llei, no era res més que una manifestació, de la *Fal·làcia de la regressió* de Galton, resultant de la metodologia d'agrupament de les dades.

3.2 Hotelling respon

En contrast amb les crítiques positives que va rebre el treball de Secrist, destaca la dura crítica de Harold Hotelling publicada al *Journal of the American Statistical Association*(*JASA*).

Harold Hotelling (1895-1973) va ser un matemàtic i economista teòric americà nascut a Minnesota, fill d'un distribuïdor de palla. Durant la seva etapa universitària a estudiar periodisme i allà va descobrir un extraordinari talent per a les matemàtiques. Hotelling va iniciar-se en les matemàtiques pures amb la seva tesi doctoral en topologia algebraica a la universitat de Princeton. Va ser professor associat de Matemàtiques a la universitat de Stanford del 1927 fins al 1931, membre del claustre de la universitat de Columbia del 1931 fins al 1946 i professor d'estadística matemàtica a la universitat de Carolina del Nord des del 1946 fins a la seva mort.

El 1933, quan es va publicar el llibre de Secrist, Hotelling era un jove professor d'estadística a la universitat de Columbia, recent doctorat, que ja havia fet grans contribucions a l'estadística teòrica, especialment en relació amb problemes econòmics.

Gran devot de la investigació, va afirmar sobre el treball de Secrist "The labor of compilation and of direct collection of data must have been gigantic", aquest va ser el comentari més amable que va fer sobre el llibre del seu company. A partir d'aquí es va encendre una disputa de rèpliques entre els dos professors.

Hotelling ressalta que el triomf de la mediocritat observat per Secrist és, en més o menys mesura, automàtic quan s'estudia una variable que rep efectes de factors estables i factors aleatoris. Explica que les nombroses taules de Secrist no proven res més que la tendència de les ràtios a oscil·lar. A més, afegeix que els resultats presentats són matemàticament obvis i que no precisen la gran acumulació de dades presentades per demostrar-los.

La primera crítica de Hotelling era educada però ferma, amb un to més proper a la pena que a la ira ja que intenta explicar a un distingit col·lega, de la manera més amable possible, que ha malgastat deu anys de la seva vida amb aquest estudi. Evidentment, aquest cop no va ser ben rebut per un orgullós Secrist que no volia creure que allò que tant d'esforç li havia costat es pogués tombar tan fàcilment,

i més amb un argument tan aparentment trivial com la fal·làcia de la regressió, concepte que ja creia haver tingut en compte.

La resposta de Secrist no es va fer esperar, i lluny de recapacitar i acceptar la mà que Hotelling li havia brindat, es va mantenir ferm en les seves conviccions, insistint que l'efecte de regressió no tenia res a veure amb els resultats. Va retreure a Hotelling no haver prestat prou atenció al llibre i haver passat per alt aquelles parts en que es té en compte un possible efecte de la regressió.

A partir d'aquest moment les rèpliques de Hotelling van abandonar el posat amable per adoptar-ne un de més àcid i directe. En aquest sentit, va escriure en una segona rèplica, "The thesis of the book, when correctly interpreted, is essentialy trivial". Va equiparar l'elevat l'esforç de recol·lecció de dades a voler demostrar les taules de multiplicar ordenant elefants en files i columnes, i repetint el mateix procés amb molts altres tipus d'animals:

> To 'prove' such a mathematical result by a costly and prolonged numerical study of many kinds of business profit and expense ratios is analogous to proving the multiplication table by arranging elephants in rows and columns, and then doing the same for numerous other kinds of animals. The performance, though perhaps entertaining, and having certain pedagogical value, is not an important contribution either to zoology or mathematics.

Cal tenir en compte un petit gest a favor de Secrist i és que, amb una actitud prudent, pròpia de qui es troba a les portes de publicar el que es podria convertir en la seva obra culminant, Secrist va sol·licitar a 38 experts estadistes i economistes d'Amèrica i Europa que fessin una revisió del llibre prèvia a la publicació. Aquesta ajuda malauradament no va evitar que publiqués el llibre sense cap modificació en relació al parany de la fal·làcia de la regressió. Tot i així, cal destacar que de totes les afirmacions que es fan al llibre, la gran majoria són certes. La mala fortuna va ser fallar en la més transcendental. En aquest sentit Hotelling va escriure:

> When in different parts of a book there are passages from which the casual reader may obtain two different ideas of what the book is providing, and when one version of the thesis is interesting but false and the other is true but trivial, it becomes the duty of the reviewer to give warning at least against the false version.

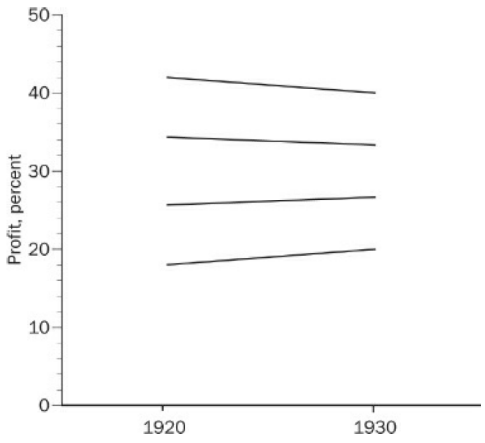
3.3 Discussió de resultats

No seria just titllar el gran treball d'investigació i recopilació de dades que va realitzar Secrist de total desastre. El llibre constava de molts arguments i afirmacions i, si bé la gran majoria eren certes, no ho era la més transcendental, el títol. Només al començar, al prefaci del llibre, podem llegir:

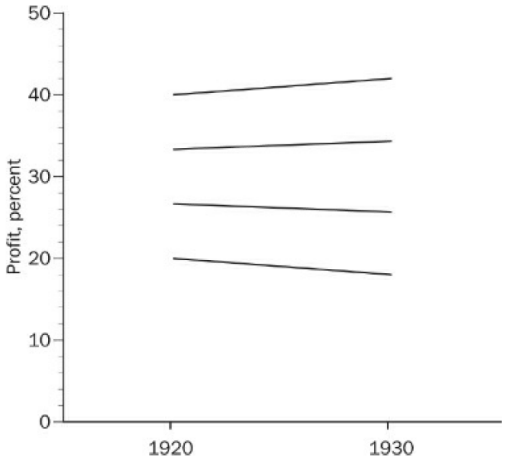
> The tendency to mediocrity in business is more than a statistical result. It is expressive of prevailing behaviour relations.

Com bé explica Hotelling en la seva primera crítica del llibre (Hotelling, 1933 [15]), de ser cert, aquest resultat seria d'immensa importància. L'errònia conclusió que sembla observar Secrist no és res més que el parany de la fal·làcia de la regressió.

La manera més simple de justificar aquesta aparent convergència en termes de regressió és argumentar que, mentre que les observacions dels marges d'un grup en una primera mesura tendeixen a desplaçar-se cap al centre (mitjana), en una segona observació, d'igual manera es pot esperar que les observacions centrals es desplacin cap als marges. Aquestes, en desplaçar-se poden moure's cap al marge més alt o cap al més baix, provocant així que les desviacions positives es cancel·lin amb les negatives i, en conseqüència, la mitjana global del grup romangui al centre. Per contra, les observacions dels grups extrems, en desplaçar-se, només poden moure's cap al centre. D'aquesta manera el mètode de seguiment de les mitjanes dels grups formats segons els beneficis del primer any d'estudi, aparenta una falsa convergència. Però, per contra, si Secrist hagués agrupat les observacions en funció dels beneficis de l'últim any, el que veuríem en fer el seguiment de les mitjanes és que divergeixen. En conseqüència, es podria demostrar estabilitat o inestabilitat segons l'interès personal, és a dir, el mètode no és adequat. Com bé mostren les gràfiques de la Figura 6, extretes de Smith (2014) [26], on s'ofereix una recreació (a petita escala) de l'estudi de Secrist.



Els quartils formats segons els beneficis del 1920 retornen a la mitjana el 1930.



Els quartils formats segons els beneficis del 1930 retornen a la mitjana el 1920.

Figura 6: *Secrist's Folly*, Gary Smith, 2014 [26].

Aquesta és la subtilesa que sovint queda oblidada. En termes de Galton equivaldria a tenir present que tan podem observar efecte de regressió dels fills sobre els pares, com dels pares sobre els fills. El que realment s'hauria de presentar per provar la suposada convergència seria un anàlisi de variància que mostrés una disminució consistent de la variància al llarg dels anys, no per a la mitjana de cada grup, sinó per a cada individu.

Per descartar que les conclusions trobades no fossin producte del mètode usat, Secrist va introduir un segon estudi basat en dades de les temperatures de 191 ciutats americanes. En ell va observar que les temperatures més altes el 1922 també eren les més altes el 1933 i viceversa per a les més baixes. Aquest resultat va permetre a Secrist concloure el fals argument, ja vist, que la mediocritat triomfava als escenaris sotmesos a forces competitives, estrictament lligades a l'esforç humà.

Arribats a aquest punt sorgeix una pregunta senzilla. Si l'efecte de regressió és un fenomen universal, per què no s'observa en les temperatures? La resposta és senzilla: sí que ho fa, i Secrist, un cop més, va fallar en veure-ho.

Veiem a continuació un exemple de gràfic de dispersió de les temperatures mitjanes del mes de gener, preses durant els anys 2016 i 2017 a la comarca del Segrià. Cap de les 12 estacions meteorològiques està a més d'una hora en cotxe d'una altra. (Veure annex 3).

Com es pot observar a la Figura 7, sí que s'observa efecte de regressió al gràfic. Al contrari de l'afirmació de Secrist, que manté que no hi ha convergència en dades meteorològiques, aquí podem veure com existeix certa variabilitat entre les ciutats amb temperatures més altes i baixes del 2016 i el 2017.

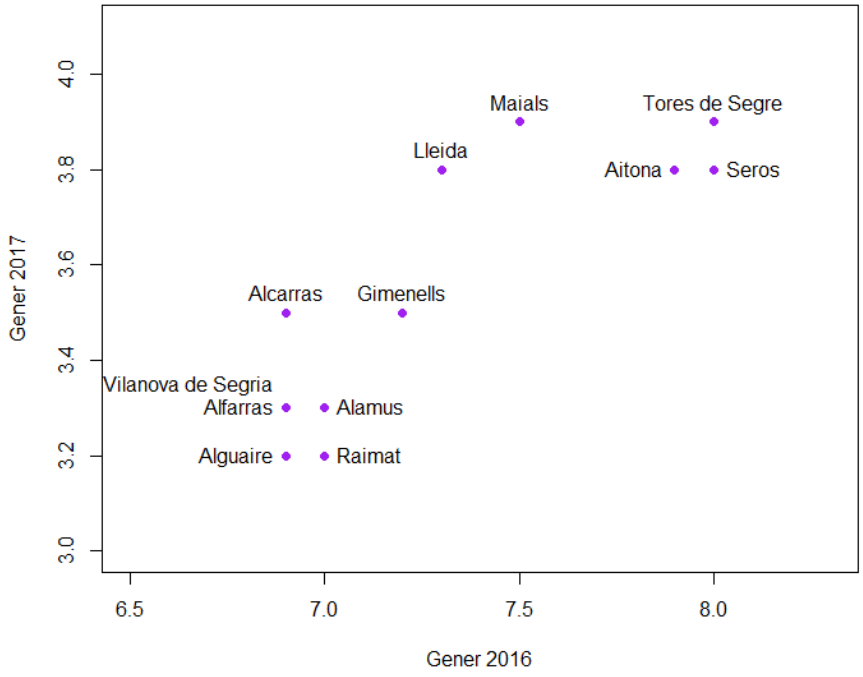


Figura 7: Diagrama de dispersió de la temperatura mitjana (en graus centígrads) del mes de gener de 2016 i 2017 de les 12 ciutats de la comarca de Segrià.

Aleshores la gran pregunta és, per què no va veure-ho Secrist? L'error, com es pot intuir a partir de l'exemple, recau en la recollida de les dades. Les seves 191

ciutats estaven seleccionades arreu dels Estats Units, i per tant, les diferències de temperatures entre elles responen a les característiques climàtiques de cada localització, de tal manera que l'efecte de regressió queda amagat. Si per el contrari, les temperatures s'haguessin pres dins d'un radi molt menor, on el clima de la zona fos més homogeni, no tindríem cap problema en observar efecte de regressió. En paraules de Hotelling "This means merely that cities do not move about".

Continuant amb l'exemple de les temperatures de Catalunya, si ara observem les dades de 8 ciutats d'arreu de Catalunya, veiem que ara sí que les temperatures més fredes i més caloroses de 2016 també ho són al 2017.

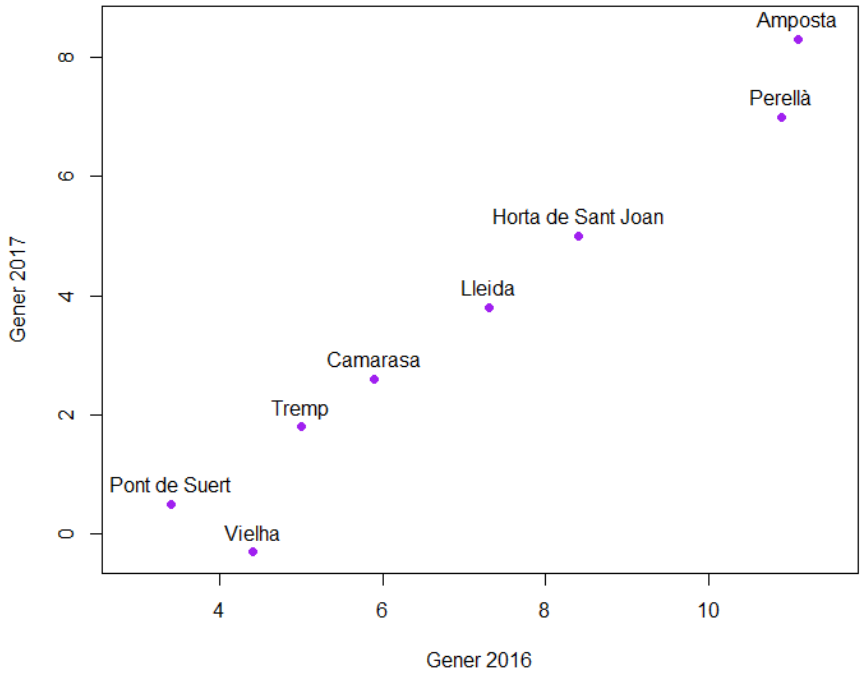


Figura 8: Diagrama de dispersió de la temperatura mitjana (en graus centígrads) del mes de gener de 2016 i 2017 de 8 ciutats preses arreu de Catalunya.

En aquesta nova gràfica que es mostra a la Figura 8, la regressió a la mediocritat queda molt més amagada en les dades.

3.4 El fenomen segueix sense entendre's

Arribats a aquest punt del treball, hom podria pensar que l'amonestació que va rebre Secrist per part de Hotelling hauria creat un precedent dins de la comunitat matemàtica que hauria posat en estat d'alerta a futures generacions. Res més lluny de la realitat. Molts són els casos que van tornar a passar per alt el fenomen, i els pocs que s'atrevien a parlar-ne als seu estudis, tenien tanta por de caure a la

trampa que aplicaven "correccionsïnnecessàries als suposats "efectes de regressióï en conseqüència generaven conclusions errònies.

Fins i tot avui en dia seguim trobant publicacions d'articles científics amb resultats erronis per culpa de la fal·làcia de la regressió, els quals arriben a unes conclusions d'allò més extraordinàries i sorprenents.

Un bon exemple d'un article actual mal interpretat seria el que va publicar l'agost de 2017 la revista *Journal of Women & Aging* titulat "Strong, healthy, energized: Striving for a health weight in an older lesbian population" per Tomisek et al. [34], i la seva corresponent resenya del desembre de 2018 feta per Halliday et al. [13].

En aquest article els autors destaquen els primers resultats sobre el programa "Strong. Healty. Energized"(SHE). Aquest programa ataca el tòpic recurrent de dissenyar un "pla saludable d'intervenció de pes"de 12 setmanes específicament per a una població de dones grans, lesbianes i amb sobrepès o obesitat. L'estudi es realitza amb una mostra de 39 participants d'entre 57 i 89 anys de Nova York. El programa SHE es marca com a objectius

(a) disminuir el contorn mitjà de la cintura en un 5%;

(b) incrementar el número de passos mitjans per dia en 2000;

(c) disminuir el consum diari de begudes edulcorades de les participants en un 25%;

(d) incrementar el percentatge de participants que compleixen amb les pautes de consum de fruita i verdures en un 25%.

Al llarg de l'estudi es comparen les observacions de les participants, referents als 4 objectius marcats, mesurades 1 setmana abans de l'inici del programa SHE amb les dades finals, mesurades 1 setmana després de finalitzar l'estudi. Un cop feta aquesta comparativa els autors de l'article conclouen que l'èxit del programa SHE és notable i que, en conseqüència, el programa hauria de ser adoptat pels col·lectius de dones de la tercera edat i LGBT. Concretament, els autors basen les seves conclusions en els resultats presentats sobre el segon objectiu. És el "marcat increment dels passos"a la categoria amb una marca basal més baixa, el que dóna suport a l'argumentació presentada pels autors de l'estudi.

Tot i així, i com remarquen Halliday et al. en la seva ressenya, l'anàlisi de l'article no porta a aquesta conclusió. Aquest increment del nombre de passos és degut a la regressió a la mitjana i no pas a una possible efectivitat del tractament. Caure en la fal·làcia de la regressió és un error recurrent en moltes àrees d'estudi, i no se'n lliuren les ciències de la salut. El principal error de disseny de l'estudi és la manca d'un grup de control amb el que poder comparar i contrastar els resultats. Negligir la regressió a la mitjana porta als autors a concloure que una certa intervenció sobre el pes pot ser efectiva quan no ho és.

Concretament a l'article es mostra la següent taula, sobre la qual basen l'afirmació que "The SHE programm was most effective for participants with low levels of physical activity and steps."

Table 8. Changes in daily steps for SHE participants ($N = 37$).

Step group tertile	Average steps at baseline	Change in average steps per day	Change in steps (maximum)
Lowest tertile	3,901.15	916.16	5,539.19
Mid tertile	6,317.37	−353.75	3,533.21
Highest tertile	9,765.73	91.25	2,582.49

Figura 9: Taula 8 de l'article [34].

Com podem observar a la taula, s'ha dividit les participants en 3 grups en funció dels nombres de passos registrats a l'inici del programa, i els individus amb valors basals per sota del terç inferior (uns aproximadament 3.900 passos al dia) obtenen valors al voltant dels 5.540 passos al dia al final del programa SHE. En l'estudi no es va incloure una aproximació analítica dels valors esperats per avaluar el canvi del nombre de passos de tota la mostra, només trobem la confirmació dels autors de no assolir l'objectiu d'incrementar en 2000 passos els passos totals al dia.

D'altra banda, pel que fa al primer objectiu (la disminució del contorn de la cintura i el pes corporal) els autors no van observar diferències significatives entre els valors basals pre estudi i els valors finals post estudi. Tot i així, conclouen que les participants que més pes han perdut al finalitzar l'estudi són aquelles que tenien les mesures més altes a l'inici. Comenten que aquest resultat no és sorprenent, com no ho hauria de ser degut a la regressió a la mitjana. El que sorprèn és que a l'estudi s'admet que els resultats referents a la pèrdua de pes són esperats i no s'apliqui la mateixa lògica pels resultats obtinguts per al nombre de passes.

Aleshores, la conclusió que hi ha prou evidència per assumir l'efectivitat del programa SHE no esta justificada, donat que els resultats podrien ser fàcilment atribuïts a l'efecte de regressió. Tot i així, això tampoc implica la no efectivitat del programa, simplement es vol posar de manifest la falta de justificació científica en la que es basen els autors del treball per assumir l'efectivitat del seu programa.

Malgrat la gran importància que té el ple coneixement i comprensió d'aquest fenomen, el que revelen molts assajos clínics és que molts investigadors no assimilen plenament aquest concepte i, per tant, no són capaços d'interpretar correctament els resultats que obtenen. Com podem prevenir o esmenar aquest fenomen? Pel que fa als assajos clínics, el més efectiu és realitzar els estudis amb un grup de control que pugui ajudar-nos a interpretar bé els resultats obtinguts. Com? Tant el grup de control com el grup d'estudi estaran sotmesos a l'efecte de regressió que, per tant, desapareixerà al comparar ambdós grups entre ells. Degut a la falta de comprensió per part dels clínics d'aquests fenòmens i metodologies estadístiques, trobem que molts investigadors no entenen per què han dut a terme un assaig clínic amb dos grups diferents, el de control i el d'estudi. En conseqüència, omplen pàgines i pàgines descrivint les respostes d'ambdós grups sense saber que no té cap rellevància. El que haurien d'extreure els clínics de treballar amb grups de control és el pensament comparatiu. Els assajos clínics controlats es fonamenten sobre la comparació, per tant, s'hauria de discutir sobre "contrast de tractament", és a dir,

diferències entre tractaments, enlloc de valorar cada tractament per si sol.

3.5 La simulació

Amb la finalitat de veure tots els arguments que s'han discutit en aquest punt és interessant presentar una simulació del que passa quan treballem en estudis que poden estar subjectes a l'efecte de regressió. Amb l'ajuda del software estadístic `R`, i el codi que es pot trobar adjunt al punt 4 de l'annex, presentarem la següent simulació.

Veurem el cas en que X i Y són dues variables aleatòries distribuïdes segons una llei Normal, és a dir, $X \sim N(\mu_x, \sigma_x^2)$ i $Y \sim N(\mu_y, \sigma_y^2)$, generant així un vector aleatori (X, Y) normal bivariant.

Per realitzar la simulació hem generat una mostra aleatòria de n=1000 observacions, amb matriu de variàncies-covariàncies teòrica,

$$\Sigma = \begin{pmatrix} 10 & 3 \\ 3 & 2 \end{pmatrix},$$

de tal manera que la variància de X és $\sigma_x^2 = 10$, la variància de Y és $\sigma_y^2 = 2$, la seva covariància és $\sigma_{xy}^2 = 3$ i les seves esperances $\mu_x = \mu_y = 0$, aleshores, $X \sim N(0, 10)$ i $Y \sim N(0, 2)$.

Amb aquestes dades hem generat el corresponent diagrama de dispersió i les dues rectes de regressió, Y sobre X i X sobre Y.

Amb la finalitat de veure les fluctuacions estadístiques que es produeixen en una regressió lineal, farem un seguiment de les observacions extremes i les observacions centrals per a les observacions de X i de Y. És a dir, tant per X com per Y, seleccionarem les observacions més extremes, enteses com les pertanyents al percentil 2 i per sobre del percentil 98, i les observacions més centrals, enteses com les compreses entre els percentils 49 i 51, i veurem a quin percentil pertany la seva observació parella. Realitzem aquest procés de manera doble per ressaltar la subtilesa del fenomen de la regressió a la mediocritat el qual sovint passa per alt als estudis i és que, tant es produeix efecte de regressió de Y sobre X com de X sobre Y.

- En primer lloc veiem la regressió de Y sobre X:

Donat que tenim n=1000 observacions, els 20 primers valors pertanyents al percentil 2, amb les seves respectives observacions per les Y són:

X	-11.46	-10.59	-9.58	-9.23	-8.75	-8.24	-7.78	-7.77	-7.35	-7.13
Y	-2.77	-2.45	-3.05	-2.04	-3.18	-2.93	-1.56	-2.64	-3.24	-1.14
q	2.2	2.9	1.3	7.1	1.0	1.9	13.5	2.3	0.8	18.6

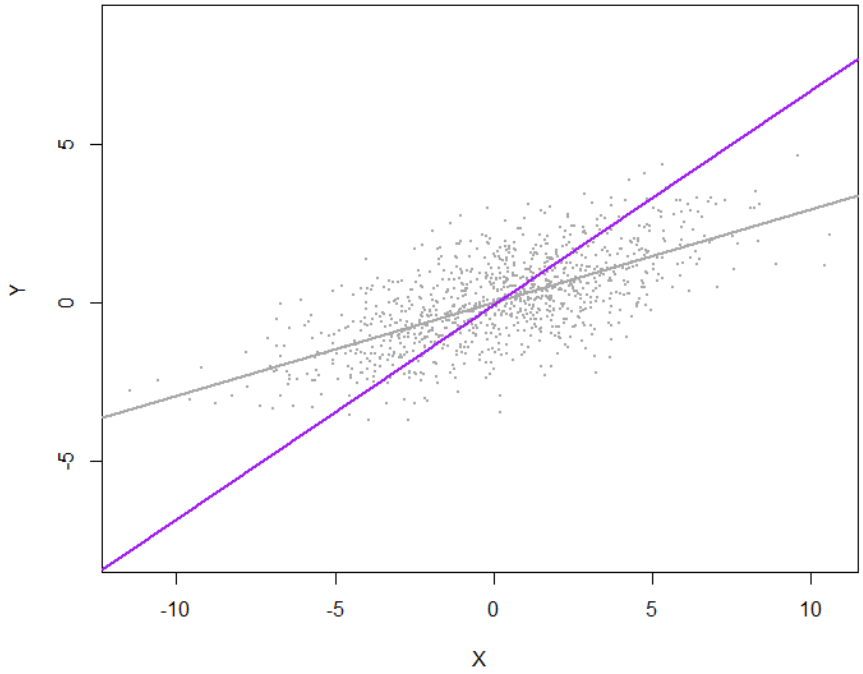


Figura 10: En gris, recta de regressió de Y sobre X, $y = -0.008 + 0.31x$. En lila, recta de regressió de X sobre Y, $y = -0.023 + 0.69x$.

-7.02	-6.94	-6.93	-6.85	-6.85	-6.83	-6.73	-6.71	-6.7	-6.58
-2.02	-3.33	-2.17	-0.52	-2.11	-1.36	-0.02	-1.67	-2.94	-1.36
7.3	0.5	6.2	33.6	6.6	16.0	47.7	10.4	1.7	16.1

L'última fila de la taula mostra per a cada x_i, $i = 1, \ldots, 20$, del percentil 2, ordenades de menor a major, el percentil al qual pertany la seva y_i. Com podem observar per a les observacions extremes de X, la seva distribució en Y és generalment menys extrema pertanyent a percentils més centrals. Dels 20 valors que observem, 14 no mantenen l'observació en Y dins del percentil 2.

Repetim ara el mateix estudi per a les 20 observacions que es troben per sobre el percentil 98:

X	6.66	6.76	6.81	6.87	6.89	7.02	7.04	7.28	7.53	7.56
Y	2.12	3.19	1.9	3.3	1.97	3.1	0.48	3.23	2.1	1.1
q	92.2	98.5	90.0	99.1	91.0	98.0	61.3	98.7	91.9	76.1

7.7	8.12	8.23	8.27	8.33	8.38	8.91	9.6	10.43	10.61
1.86	2.97	2.98	3.5	1.95	3.11	1.22	4.62	1.2	2.12
89.2	97.4	97.5	99.5	90.7	98.2	79.5	100.0	78.7	92.3

Igual que abans, hem obtingut el resultat esperat. De les 20 observacions més extremes de X, 13 observacions no mantenen aquesta condició per a la parella y_i i pertanyen a percentils inferiors al 98.

Finalment, veiem el cas en que les observacions són centrals:

X	0.261	0.266	0.282	0.284	0.317	0.326	0.330	0.332	0.332	0.344
Y	-1.664	0.108	-0.288	1.352	-0.274	0.622	0.507	-0.257	0.792	0.179
q	10.7	51.0	39.7	81.0	40.1	65.3	62.3	41.0	70.3	53.6

0.347	0.363	0.365	0.368	0.373	0.379	0.382	0.388	0.389	0.396
0.653	0.65	0.072	-0.11	1.766	0.24	1.94	2.6	1.201	-0.791
66.3	65.9	50.0	45.5	87.8	54.6	90.6	95.4	79.2	26.6

Veiem que per a aquestes observacions els resultats són encara més marcats, només 2 observacions conserven la seva parella y_i en percentils centrals.

- En segon lloc veiem la regressió de X sobre Y:

A continuació repetirem el mateix procés realitzat anteriorment però partint de la distribució de Y i amb l'objectiu d'esperar els mateixos resultats.

Primer de tot, estudiem els percentils extrems 2 i 98, respectivament:

Y	-3.71	-3.68	-3.52	-3.47	-3.33	-3.29	-3.25	-3.24	-3.18	-3.18
X	-3.92	-2.71	-4.55	0.2	-6.94	-5.73	-6.32	-7.35	-2.51	-8.75
q	9.3	17.7	6.0	48.2	1.2	3.3	2.6	0.9	19.4	0.5

-3.16	-3.1	-3.05	-3.05	-3.01	-2.94	-2.94	-2.94	-2.93	-2.90
-3.36	-2.09	-9.58	-2.80	-2.15	-4.53	-6.7	0.21	-8.24	-5.44
12.3	23.5	0.3	16.7	23.3	6.1	1.9	48.6	0.6	4.1

Hem obtingut que de les 20 observacions de Y pertanyents al percentil 2, només 6 no tenen una tendència a apropar-se a la mitjana i per tant, segueixen formant part del percentil 2 de les observacions de les X.

Y	3.10	3.11	3.12	3.16	3.19	3.22	3.23	3.24	3.24	3.27
X	6.14	8.38	1.10	3.09	6.76	5.89	7.28	5.75	4.16	4.85
q	97.0	99.6	59.7	82.5	98.2	96.1	98.8	95.8	89.3	93.3

3.3	3.31	3.32	3.41	3.5	3.63	3.88	4.06	4.37	4.62
6.86	3.69	6.40	2.87	8.27	5.42	4.29	4.74	5.34	9.6
98.4	86.6	97.6	80.7	99.4	95.5	90.2	92.4	95.1	99.8

Una vegada més, per a les 20 observacions més extremes per sobre del percentil 98 de la distribució de Y, 14 tendeixen a retrocedir cap a la mitjana global i, per tant, a percentils inferiors que el 98, en la seva distribució de X.

Finalment, veiem que aquesta tendència torna a repetir-se per a les observacions centrals de la distribució de Y:

Y	0.038	0.038	0.039	0.045	0.057	0.059	0.061	0.068	0.069	0.072
X	0.981	3.485	-3.128	0.858	-2.646	-1.787	-0.312	-1.922	0.181	0.365
q	58.3	85.1	14.2	56.3	18.3	26.2	41.4	25.2	47.7	50.3

0.075	0.084	0.089	0.091	0.093	0.094	0.098	0.102	0.105	0.108
-3.08	0.127	1.088	0.703	1.769	-0.199	0.004	-6.065 1.308	0.266	
14.6	46.8	59.5	55.0	68.9	42.7	45.2	2.8	62.7	49.2

Un altre cop tornem a veure que la tendència general de les observacions (18 de les 20) és que les seves parelles es desplacin cap als extrems enlloc de romandre a percentils centrals.

4 Relació amb Shrinkage

4.1 La paradoxa d'Stein

L'objectiu essencial de l'estadística inferencial és obtenir informació de grans conjunts de dades o poblacions a partir de subconjunts anomenats *mostres*. Per poder obtenir aquesta informació s'ha de resoldre el problema fonamental de l'estimació dels paràmetres d'una determinada distribució associada a la característica d'estudi. Per fer això, donat un conjunt d'observacions $X = \{x_1, \ldots, x_k\}$ s'utilitzen els denominats *estimadors* $U(X) \equiv U(x_1, \ldots, x_k)$, els valors dels quals han de ser "propers" al paràmetre θ desconegut.

Aquests estimadors tenen diferents característiques que fan que siguin considerats més bons o menys. Per valorar la bondat d'un estimador podem estudiar-ne diferents propietats.

Diem que un estimador és més *eficient* que un altre si la seva variància és menor, és a dir, siguin U i V dos estimadors diferents de θ tals que

$$\text{var}(V) < \text{var}(U),$$

aleshores diem que V és més eficient que U.

A l'hora de jutjar el bon comportament d'un estimador podem definir la *funció de pèrdua quadràtica* en l'espai de paràmetres $\Theta \subset \mathbb{R}^d$, $d \geq 1$, per mesurar com de lluny està el valor $U(X)$ del paràmetre θ. Ens proporciona una idea del *cost* de donar $U(X)$ com a estimació d'un valor concret $\theta \in \Theta$.

$$\begin{aligned} L : \Theta \times \Theta &\longrightarrow \mathbb{R}^+ \\ (\theta_1, \theta_2) &\longmapsto \sum_{i=1}^{k} (\theta_1 - \theta_2)^2, \end{aligned}$$

per $\theta_1, \theta_2 \in \Theta$.

En un model estadístic paramètric, amb espai mostral $\mathscr{X}$ i espai de paràmetres Θ, anomenem *espai d'estimadors de quadrat integrable* a l'espai d'estimadors $U(X)$ de $\theta \in \Theta$

$$\mathscr{D} = \{U \,|\, U : \mathscr{X} \to \Theta\}$$

tals que

$$\mathbb{E}_\theta(||U(X)||^2) < \infty,$$

per tot $\theta \in \Theta$.

Per tal de tenir una idea global del *cost* de U considerem la *funció de risc* (o *funció de pèrdua esperada*) definida com l'esperança matemàtica de la funció de pèrdua. Aleshores per a un estimador $U \in \mathscr{D}$ del paràmetre $\theta \in \Theta$ es defineix com

$$R_U(\theta) \equiv R(U, \theta) = \mathbb{E}_\theta[L(U(X), \theta)]$$

definida a $\mathbb{R}^+$. D'ara en endavant, quan no hi hagi ambigüitat, ometrem el subíndex θ per tal de simplificar la notació.

Finalment, diem que un estimador $U \in \mathscr{D}$ és *inadmissible* si existeix un altre estimador $V \in \mathscr{D}$ tal que el seu risc sigui uniformement menor,

$$R_V(\theta) \leq R_U(\theta),\ \forall \theta \in \Theta.$$

Si no existeix cap V aleshores es diu que U és *admissible*. Quan la desigualtat és estricta, es diu que V *domina* U.

4.1.1 L'estimador James-Stein

Una part important de l'anàlisi de variància s'ocupa d'estudiar l'estimació simultània d'una col·lecció de mitjanes normals. En una població normal la mitjana $\overline{x}$ és un estimador admissible de μ, és a dir, no existeix cap altre estimador uniformement millor. Aquest resultat és igualment vàlid quan els nombre de mitjanes a estimar són dues.

L'enlluernador descobriment el 1956 de Charles Stein (1920 - 2016), i millorat el 1961, en col·laboració amb el seu estudiant Willard James, va ser provar que aquest estimador ordinari no és admissible per a un vector de mitjanes de dimensió $k \geq 3$. La demostració del teorema és constructiva, és a dir, proporciona un nou estimador per a $k > 2$ mitjanes amb risc uniformement inferior a l'estimador de màxima versemblança obtingut a partir de les mitjanes individuals. Més formalment tenim que:

Teorema. *Siguin $X_1, X_2, \ldots, X_k$ una col·lecció de variables independents normalment distribuïdes, i amb igual variància σ^2, $X_i \sim N(\theta_i, \sigma^2)$, $i = 1, \ldots, k$. Sigui $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$ un vector de paràmetres desconegut. Aleshores, l'estimació ordinària per a cada θ_i, $\hat{\theta}_i = X_i$, no és admissible per a $k \geq 3$, amb la funció de pèrdua quadràtica.*

Stein presenta l'estimador de θ donat per

$$\hat{\theta}^{JS} = \left(1 - \frac{k-2}{S^2}\right) X,$$

on $S^2 = \sum_{j=1}^{k} X_j^2$. Aquest estimador és "millor"en el sentit que té uniformement menys *risc* per θ, és a dir, en termes de la *funció de pèrdua quadràtica*,

$$\sum_{i=1}^{k} (\theta_i - \hat{\theta}_i^{JS})^2 \leq \sum_{i=1}^{k} (\theta_i - X_i)^2.$$

L'estimador de James-Stein pot ser considerat com una mitjana ponderada de 0 i X, per aquest motiu també s'anomena estimador de contracció (*Shrinkage*), per

ser una contracció de l'estimador ordinari $\hat{\theta} = X$ cap al 0, tot i que si $S^2 < k-2$ obtenim valors negatius.

Demostració. Per veure que l'estimador James-Stein és un estimador admissible davant de l'estimador ordinari, hem de veure que:

$$R(\hat{\theta}^{JS}) < R(X).$$

Com que $R(X) = \sum_{i=1}^{k} R(X_i)$, anem a calcular en detall el risc de cada component i-èssima. Donat que $X_i \sim N(\theta_i, \sigma^2)$, tenim que

$$R(X) = \mathbb{E}\sum_{i=1}^{k}(\theta_i - X_i)^2 = \sum_{i=1}^{k}\mathbb{E}(\theta_i - X_i)^2 = k\sigma^2.$$

D'altra banda, hem de calcular $R(\hat{\theta}^{JS})$:

$$\begin{aligned}
R(\hat{\theta}_i^{JS}) &= \mathbb{E}\sum_{i=1}^{k}(\theta_i - \hat{\theta}_i^{JS})^2 = \sum_{i=1}^{k}\mathbb{E}\left[\theta_i - \left(1 - \frac{k-2}{\sum_{j=1}^{k} X_j^2}\right) X_i\right]^2 = \\
&= \sum_{i=1}^{k}\mathbb{E}\left[(\theta_i - X_i) + \frac{k-2}{\sum_{j=1}^{k} X_j^2} X_i\right]^2 = \\
&= \sum_{i=1}^{k}\mathbb{E}\left[(\theta_i - X_i)^2 + \left(\frac{k-2}{\sum_{j=1}^{k} X_j^2} X_i\right)^2 + 2(\theta_i - X_i)\frac{k-2}{\sum_{j=1}^{k} X_j^2} X_i\right] = \\
&= \underbrace{\sum_{i=1}^{k}\mathbb{E}(\theta_i - X_i)^2}_{\text{A}} + \underbrace{\sum_{i=1}^{k}\mathbb{E}\left[\frac{k-2}{\sum_{j=1}^{k} X_j^2} X_i\right]^2}_{\text{B}} + \underbrace{\sum_{i=1}^{k}\mathbb{E}\left[2(\theta_i - X_i)\frac{k-2}{\sum_{j=1}^{k} X_j^2} X_i\right]}_{\text{C}}.
\end{aligned}$$

Desenvolupem els tres sumands A, B i C per separat:

El primer terme,

$$A = \sum_{i=1}^{k}\mathbb{E}(\theta_i - X_i)^2 = k\sigma^2.$$

El segon terme,

$$\begin{aligned}
B = \sum_{i=1}^{k}\mathbb{E}\left[\frac{k-2}{\sum_{j=1}^{k} X_j^2} X_i\right]^2 = \mathbb{E}\sum_{i=1}^{k}\left[\frac{(k-2)^2}{(\sum_{j=1}^{k} X_j^2)^2} X_i^2\right] = (k-2)^2\mathbb{E}\left[\frac{\sum_{i=1}^{k} X_i^2}{(\sum_{j=1}^{k} X_j^2)^2}\right] = \\
= (k-2)^2\mathbb{E}\left[\frac{1}{\sum_{j=1}^{k} X_j^2}\right].
\end{aligned}$$

Per últim, per calcular l'esperança del tercer terme, veiem el següent Lema de Stein.

Lema. *Sigui $X \sim N(\mu, \sigma^2)$ i $g : \mathbb{R} \longrightarrow \mathbb{R}$ diferenciable, tal que $\mathbb{E}|g'(X)| < +\infty$. Aleshores, es verifica que,*

$$\mathbb{E}\{(X-\mu)g(X)\} = \sigma^2\,\mathbb{E}(g'(X)).$$

Demostració.

$$\mathbb{E}\{(X-\mu)g(X)\} = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{+\infty} g(x)(x-\mu)e^{\frac{-(x-\mu)^2}{2\sigma^2}}\,dx = (\star)$$

Integrant per parts:

$$\begin{aligned} u &= g(x), & du &= g'(x)\,dx,\\ dv &= (x-\mu)e^{\frac{-(x-\mu)^2}{2\sigma^2}}\,dx, & v &= -\sigma^2 e^{\frac{-(x-\mu)^2}{2\sigma^2}}, \end{aligned}$$

$$(\star) = \left[\frac{1}{\sigma\sqrt{2\pi}}\,g(x)\,\sigma^2\,e^{\frac{-(x-\mu)^2}{2\sigma^2}}\right]_{-\infty}^{+\infty} + \quad \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{+\infty}\sigma^2\,e^{\frac{-(x-\mu)^2}{2\sigma^2}}\,g'(x)\,dx =$$

$$= \sigma^2\left[\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{+\infty} g'(x)\,e^{\frac{-(x-\mu)^2}{2\sigma^2}}\,dx\right] = \sigma^2\mathbb{E}(g'(x)).$$

Seguint amb la demostració del Teorema, usem el Lema de Stein a la segona igualtat,

$$C = \sum_{i=1}^{k}\mathbb{E}\left[2(\theta_i - X_i)\frac{k-2}{\sum_{j=1}^k X_j^2}X_i\right] = -2(k-2)\sum_{i=1}^{k}\mathbb{E}\left[(X_i-\theta_i)\frac{X_i}{\sum_{j=1}^k X_j^2}\right] =$$
$$= -2(k-2)\sigma^2\sum_{i=1}^{k}\mathbb{E}\left[\frac{\partial}{\partial X_i}\frac{X_i}{\sum_{j=1}^k X_j^2}\right] = -2(k-2)\sigma^2\sum_{i=1}^{k}\mathbb{E}\left[\frac{\sum_{j=1}^k X_j^2 - 2X_i^2}{(\sum_{j=1}^k X_j^2)^2}\right] =$$
$$= -2(k-2)\sigma^2\mathbb{E}\sum_{i=1}^{k}\left[\frac{\sum_{j=1}^k X_j^2}{(\sum_{j=1}^k X_j^2)^2} - \frac{2X_i^2}{(\sum_{j=1}^k X_j^2)^2}\right] =$$
$$= -2(k-2)\sigma^2\mathbb{E}\left[\sum_{i=1}^{k}\frac{1}{\sum_{j=1}^k X_j^2} - \frac{2\sum_{i=1}^k X_i^2}{(\sum_{j=1}^k X_j^2)^2}\right] =$$
$$= -2(k-2)\sigma^2\mathbb{E}\left[\frac{k}{\sum_{j=1}^k X_j^2} - \frac{2}{\sum_{j=1}^k X_j^2}\right] =$$
$$= -2(k-2)^2\sigma^2\mathbb{E}\left[\frac{1}{\sum_{j=1}^k X_j^2}\right].$$

Finalment, reunint els termes A, B i C tenim que,

$$R(\hat{\theta}^{JS}) = A + B + C = k\sigma^2 + (k-2)^2\,\mathbb{E}\left[\frac{1}{\sum_{j=1}^{k} X_j^2}\right] - 2(k-2)^2\sigma^2\,\mathbb{E}\left[\frac{1}{\sum_{j=1}^{k} X_j^2}\right] =$$

$$= k\sigma^2 - \underbrace{(k-2)^2\sigma^2\,\mathbb{E}\left[\frac{1}{\sum_{j=1}^{k} X_j^2}\right]}_{>0} < k\sigma^2 = R(X),$$

per tot $k > 2$.

Admissibilitat de l'estimador James-Stein

L'estimador James-Stein, $\hat{\theta}^{JS}$, és un estimador de la família d'estimadors $\hat{\theta}^c$,

$$\hat{\theta}^c = \left(1 - \frac{c}{S^2}\right) X,$$

on $c \in \mathbb{R}$ és una constant que depèn de la dimensió de la mostra, $0 < c < 2(k-2)$. Cada estimador $\hat{\theta}^c$ és un millor estimador que l'ordinari i pren el seu valor òptim quan $c = k - 2$, el cas de James-Stein.

Tot i així, existeixen estimadors uniformement millors que el de James-Stein, com l'estimador de la part positiva,

$$\hat{\theta}^+ = \left(1 - \frac{k-2}{S^2}\right)^+ X,$$

on $(\cdot)^+$ és la funció *part positiva*, $(z)^+ = \max\{0, z\}$. A més, aquest estimador corregeix el comportament erràtic de $\hat{\theta}^{JS}$ quan $x \to 0$, en què $\hat{\theta}^{JS} \to \pm\infty$. (Cassella-Berger, 2002 [1]).

4.1.2 L'estimador Efron-Morris

Des de la seva aparició, diverses han estat les variacions de, $\hat{\theta}^{JS}$, que s'han proposat. Una d'elles és l'*estimador Efron-Morris*, proposada el 1975 per Bradley Efron i Carl Morris. A diferència de l'estimador James-Stein, aquest estimador contrau cada X_i cap a la mitjana global $\overline{X} = \frac{1}{k}\sum_{i=1}^{k} X_i$ i domina l'estimador ordinari per a $k \geq 4$. Aquest estimador es defineix dins de la família d'estimadors

$$\hat{\theta}_i^d = \overline{X} + d(X_i - \overline{X}),$$

$$\text{on } \; d = 1 - \frac{c}{S'^2} \quad \text{amb} \quad 0 < c < 2(k-3), \;\; \text{constant}, \quad \text{i} \quad S'^2 = \sum_{i=1}^{k} (X_i - \overline{X})^2,$$

de tal manera que, l'estimador Efron-Morris pren el valor òptim per $c = k - 3$ dins de la família d'estimadors $\hat{\theta}^d$,

$$\hat{\theta}^{EM} = \overline{X} + \left(1 - \frac{k-3}{S'^2}\right)(X_i - \overline{X})^2.$$

4.2 L'estimador d'Stein com un problema de regressió

La complexitat d'entendre la paradoxa de Stein recau principalment en el fet que la seva demostració, generalment, depèn del càlcul explícit del risc de l'estimador sense aportar cap visió intuïtiva del perquè d'aquest fenomen. Moltes vegades el símil que s'utilitza a l'hora d'introduir la paradoxa és: *"Com pot, informació sobre el preu de les pomes a Washington i sobre el preu de les taronges a Florida, ser usat per millorar l'estimació del preu del vi francès?"* (Stigler, 1990 [31]).

A banda de la demostració rigorosa del teorema, són diverses les justificacions que es poden trobar de la paradoxa. En destaquen l'argument geomètric (Stein, 1961 [28]), l'argument de l'estimador Bayes empíric (Efron-Bradley, 2010 [5]), l'argument freqüentista (A. K. Gupta i E. A. Peña, 1991 [12]) i també l'argument "Galtonià"(Stigler, 1990 [31]).

De tots ells, el que destaquem en aquest treball és l'argument Galtonià, basat en una regressió a la mitjana presentat per Stigler. En aquest article ens presenta una construcció més intuïtiva de la paradoxa, presentant-la com un problema d'ajust.

Stigler parteix de la forma més simple de la paradoxa en que la variància del vector de mitjanes és, per totes, igual a 1. La situació és aquesta:

Problema. Sigui una col·lecció de mesures independents $X_1, X_2, \ldots, X_k$ distribuïdes amb una llei de probabilitat normal $N(\theta_i, 1)$, cada una mesurant un paràmetre desconegut θ_i diferent. Es volen estimar els paràmetres θ_i amb funció de pèrdua quadràtica $L(\theta, \hat{\theta}) = \sum_{i=1}^{k}(\theta - \hat{\theta})^2$ i jutjar la seva bondat amb la funció del risc $R(\theta, \hat{\theta}) = \mathbb{E}L(\theta, \hat{\theta})$.

Solució. Per tractar la paradoxa com un problema d'ajust hem de treballar amb els parells $(X_i, \theta_i),\ i = 1, \ldots, k$ on l'element X_i és conegut i l'element θ_i és desconegut. En conseqüència, els punts (X_i, θ_i) no és poden visualitzar en un gràfic. Tot i així, seria molt útil si poguéssim imaginar-nos quina aparença tindria un gràfic com aquest. En aquest sentit Stigler presenta una gràfica hipotètica (Figura 11). Així doncs, com que les observacions X_i es comporten segons una $N(\theta_i, 1)$, podem pensar les X_i com un desplaçament de cada θ_i segons una $N(0, 1)$. D'aquesta manera, les desviacions horitzontals dels punts de la recta $\theta = X$ són independents $N(0, 1)$. A més, com $\mathbb{E}(\overline{X}) = \overline{\theta} = \frac{1}{k}\sum_{i=1}^{k}\theta_i$ i $\mathrm{Var}(\overline{X}) = \frac{1}{k}$, podem esperar que el punt $(\overline{X}, \overline{\theta})$ caigui a prop de la recta $\theta = X$.

L'objectiu que ens plantegem és: donats els X_i, estimar els θ_i sense suposar cap distribució. El plantejament que proposa Stigler és un raonament invers, és a dir, per veure per què l'estimador ordinari pot ser millorat, proposa estudiar el problema pensant què faríem si coneguéssim la distribució conjunta dels parells (X_i, θ_i). Ajustem per regressió i calculem els valors $\hat{\theta}(X) = \mathbb{E}(\theta|X)$, de la funció de regressió teòrica, i generar estimacions $\hat{\theta}$ de θ avaluant cada valor X_i. Aquest plantejament no és realista ja que no coneixem quina és la distribució de θ donat X, i en conseqüència, no podem calcular $\mathbb{E}(\theta|X)$. A més, tampoc farem cap assumpció sobre si els θ_i poden ser descrits per una funció de distribució de probabilitat.

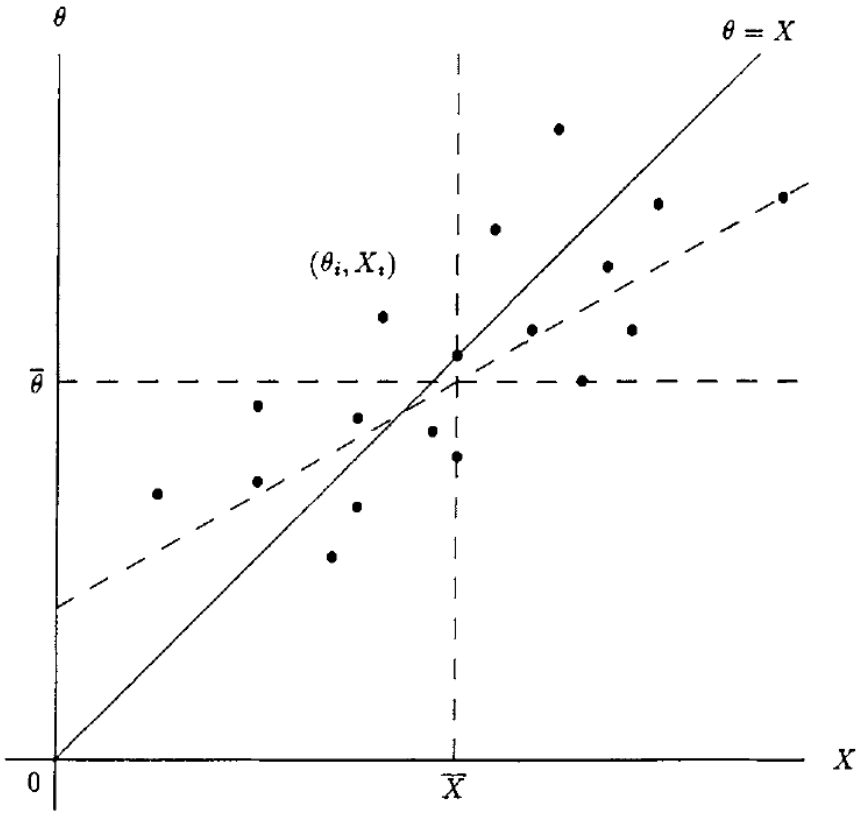


Figura 11: Diagrama de dispersió bivariant hipotètic de les θ_i contra les X_i $i = 1, \ldots, k$ (Stigler, 1990 [31]) .

Pel contrari, el que sí que coneixem és la distribució de X donat θ, $X \sim N(\theta, 1)$, i per tant, podem calcular l'altra recta de regressió $\mathbb{E}(X|\theta) = \theta$. Aquesta recta de regressió teòrica correspon a la recta $\theta = X$, que ve proporcionada pels estimadors ordinaris $\hat{\theta}_i = X_i$. D'aquesta manera, podem interpretar l'estimador ordinari com l'estimador basat en la recta de regressió "equivocada"(X sobre θ, $\mathbb{E}(X|\theta)$), en lloc de en la "bona"(θ sobre X, $\mathbb{E}(\theta|X)$). Ambdues rectes de regressió poden ser diferents, ja que en una regressió lineal de X i Y la recta de Y sobre X és més plana en relació a l'eix de les abscisses que la de X sobre Y. Aquest fet és el que, intuïtivament, suggereix que l'estimador ordinari pot ser millorat i, de quina manera: intentar estimar "$\mathbb{E}(\theta|X)$", o el sentit que se li pugui donar quan θ no segueix una distribució paramètrica.

Per tal d'encarar la cerca d'aquest estimador admissible, donat que l'estimador ordinari és un estimador lineal, podem intentar buscar un "millor estimador lineal", que minimitzi la funció de pèrdua quadràtica. Tornant al cas hipotètic, si els valors θ_i fossin coneguts, aquest "millor estimador lineal"seria el generat per l'estimador mínims quadrats ordinaris de la recta de regressió de θ sobre X. Com hem vist als punts anteriors aquesta seria,

$$\theta_i = \overline{\theta} + \hat{\beta}(X_i - \overline{X}),$$

on

$$\hat{\beta} = \frac{\sum_{i=1}^{k}(X_i - \overline{X})(\theta_i - \overline{\theta})}{\sum_{i=1}^{k}(X_i - \overline{X})^2}.$$

Si poguéssim donar una estimació de les funcions $\overline{\theta}$ i en conseqüència $\hat{\beta}$, tindríem una estimació de la regressió de θ sobre X.

Per una banda, l'estimador més obvi (uniforme, de mínima variància i no esbiaixat) de $\overline{\theta}$ és $\overline{X}$.

D'altra banda, hem de construir un estimador per $\hat{\beta}$. Donat que desconeixem els valors θ_i, no podem calcular el numerador de $\hat{\beta}$ ($\sum_{i=1}^{k}(X_i - \overline{X})(\theta_i - \overline{\theta})$), per tant, mitjançant un argument bayesià, intentarem trobar-ne una aproximació. Suposem, un altre cop hipotèticament, que aquests valors θ_i són independents distribuïts segons una distribució prior qualsevol, coneguda o no, amb moment de segon ordre finit.

Aleshores la covariància de la mostra (X, θ) es pot estimar per l'estimador no esbiaixat:

$$\operatorname{cov}(X, \theta) = \frac{\sum_{i=1}^{k}(X_i - \overline{X})(\theta_i - \overline{\theta})}{k-1},$$

i tenim que

$$\text{numerador}(\hat{\beta}) = (k-1)\operatorname{cov}(X, \theta).$$

Paral·lelament, com hem definit anteriorment al realitzar la Figura 11, $X = \theta + \varepsilon$ on $\varepsilon \sim N(0,1)$ independent de θ, tenim que la covariància de la mostra es pot estimar d'aquesta altra manera:

$$\operatorname{cov}_{bis}(X, \theta) = \operatorname{cov}(\theta + \varepsilon,\, \theta) = \operatorname{cov}(\theta, \theta) + \operatorname{cov}(\varepsilon, \theta) = \operatorname{var}(\theta) + 0 = \operatorname{var}(\theta).(\star)$$

Podem calcular $\operatorname{var}(\theta)$ a partir de $X = \theta + \varepsilon$,

$$\operatorname{var}(X) = \operatorname{var}(\theta + \varepsilon) = \operatorname{var}(\theta) + \operatorname{var}(\varepsilon)$$

$$\Downarrow$$

$$\operatorname{var}(\theta) = \operatorname{var}(X) - \operatorname{var}(\varepsilon),$$

i per tant,

$$(\star)\, \operatorname{cov}_{bis}(X, \theta) = \operatorname{var}(\theta) = \operatorname{var}(X) - \operatorname{var}(\varepsilon) = \operatorname{var}(X) - 1.$$

La variància de X la calculem a partir de la seva distribució marginal i per tant prenem l'estimador no esbiaixat

$$\operatorname{var}(x) = \frac{\sum(X_i - \overline{X})^2}{k-1},$$

aleshores,

$$\operatorname{cov}_{bis}(X, \theta) = \frac{\sum_{i=1}^{k}(X_i - \overline{X})^2}{k-1} - 1.$$

Podem comprovar que ambdues covariàncies

$$\mathrm{cov}(X,\theta)=\frac{\sum_{i=1}^k(X_i-\overline{X})(\theta_i-\overline{\theta})}{k-1}$$

i

$$\mathrm{cov}_{bis}(X,\theta)=\frac{\sum_{i=1}^k(X_i-\overline{X})^2-(k-1)}{k-1},$$

tenen la mateixa esperança.

Per a realitzar els càlculs de comprovació prescindirem del factor constant $\frac{1}{k-1}$ d'ambdues expressions. Per una banda tenim,

$$\mathbb{E}\sum_{i=1}^k(X_i-\overline{X})(\theta_i-\overline{\theta})=\mathbb{E}\sum_{i=1}^k(X_i\theta_i-X_i\overline{\theta}-\overline{X}\theta_i+\overline{X}\,\overline{\theta})=$$

$$=\sum_{i=1}^k\mathbb{E}(X_i\theta_i-X_i\overline{\theta}-\overline{X}\theta_i+\overline{X}\,\overline{\theta})=\sum_{i=1}^k\left[\theta_i\mathbb{E}(X_i)-\overline{\theta}\mathbb{E}(X_i)-\overline{X}\theta_i+\overline{X}\,\overline{\theta}\right]=$$

$$=\sum_{i=1}^k(\theta_i^2-\overline{\theta}\theta_i-\overline{\theta}\theta_i+\overline{\theta}^2)=\sum_{i=1}^k(\theta_i-\overline{\theta})^2.$$

Per altra banda,

$$\mathbb{E}\left[\sum_{i=1}^k(X_i-\overline{X})^2-(k-1)\right]=\mathbb{E}\sum_{i=1}^k(X_i^2-2X_i\overline{X}+\overline{X}^2)-(k-1)=$$

$$=\sum_{i=1}^k\mathbb{E}(X_i^2-2X_i\overline{X}+\overline{X}^2)-(k-1)\quad(\star)$$

Anem a calcular cada terme de l'esperança per separat:

$$\mathbb{E}(X_i^2)=\mathbb{E}(\theta_i+\varepsilon)^2=\mathbb{E}(\theta_i^2+2\varepsilon\theta_i+\varepsilon^2)=\mathbb{E}(\theta_i^2)+2\theta_i\mathbb{E}(\varepsilon)+\mathbb{E}(\varepsilon^2)\overset{(*)}{=}\mathbb{E}(\theta_i^2)+1=\theta_i^2+1$$

En la igualtat $(*)$ hem usat que $\varepsilon\sim N(0,1)$ i per tant, $2\theta_i\mathbb{E}(\varepsilon)=2\theta_i\cdot 0=0$ i com que $\mathrm{var}(\varepsilon)=\mathbb{E}(\varepsilon^2)+\mathbb{E}(\varepsilon)^2$, tenim que, $\mathbb{E}(\varepsilon^2)=\mathrm{var}(\varepsilon)-\mathbb{E}(\varepsilon)^2=1$.

$$\mathbb{E}(X_i\overline{X})=\frac{1}{k}\sum_{j=1}^k\mathbb{E}(X_iX_j)=\frac{1}{k}\mathbb{E}(X_i^2)+\frac{1}{k}\sum_{j\neq i}^k\mathbb{E}(X_iX_j)=\frac{1}{k}(\theta_i^2+1)+\frac{1}{k}\theta_i\sum_{j\neq i}^k\theta_j=$$

$$=\frac{1}{k}(1+\theta_ik\overline{\theta})=\frac{1}{k}+\theta_i\overline{\theta}.$$

$$\mathbb{E}(\overline{X}^2)=\frac{1}{k^2}\mathbb{E}\left(\sum_{i=1}^kX_i\right)^2=\frac{1}{k^2}\mathbb{E}\left(\sum_{i=1}^kX_i^2+\sum_{i=1}^k\sum_{j\neq 1}^kX_iX_j\right)=\frac{1}{k^2}\left(\sum_{i=1}^k(\theta_i^2+1)+\sum_{i=1}^k\sum_{j\neq i}^k\theta_i\theta_j\right)=$$

$$=\frac{1}{k^2}\left(k+\sum_{i=1}^k\sum_{j=1}^k\theta_i\theta_j\right)=\frac{1}{k^2}(k+k^2\overline{\theta}^2)=\frac{1}{k}+\overline{\theta}^2.$$

Reunint tots els termes,

$$(\star) = \sum_{i=1}^{k} \left[\theta_i^2 + 1 - 2\left(\frac{1}{k} + \theta_i \overline{\theta}\right) + \frac{1}{k} + \overline{\theta}^2 \right] - (k-1) = \sum_{i=1}^{k} (\theta_i - \overline{\theta})^2.$$

Així doncs, amb independència de la hipotètica distribució de θ, com teníem que

$$\text{numerador}(\hat{\beta}) = (k-1)\,\text{cov}(X, \theta),$$

podem prendre,

$$\text{numerador}(\hat{\beta}) = (k-1)\,\text{cov}_{bis}(X, \theta),$$

de tal manera que, com a estimació de la funció aleatòria paramètrica $\hat{\beta}$, obtenim l'expressió

$$\hat{\beta} = \frac{\sum_{i=1}^{k}(X_i - \overline{X})^2 - (k-1)}{\sum_{i=1}^{k}(X_i - \overline{X})^2} = 1 - \frac{k-1}{\sum_{i=1}^{k}(X_i - \overline{X})^2} = 1 - \frac{k-1}{S'^2}.$$

Finalment, un cop estimades les funcions $\overline{\theta}$ i $\hat{\beta}$, podem donar la forma de l'estimador generat per la recta de mínims quadrats com

$$\hat{\theta}_i^{\mathbf{C}} = \overline{X} + \left(1 - \frac{k-1}{S'^2}\right)(X_i - \overline{X}),$$

que com podem observar, és l'estimador Efron-Morris per a $c = k - 1$. No hem obtingut el valor de $c = k - 3$ que optimitza la família d'estimadors $\hat{\theta}_i^{\mathbf{C}}$ però, tot i així, té un risc uniformement menor mentre $k - 1 < 2(k-3)$, o equivalentment $k > 5$.

Per derivar d'aquest raonament l'estimador James-Stein, és suficient considerar la família d'estimadors lineals en X que no tenen terme independent, és a dir, $\hat{\theta}_i^c = b\,X_i$. Aleshores, l'estimador mínims quadrats queda de la següent forma,

$$\hat{\beta} = \frac{\sum_{i=1}^{k} \theta_i X_i}{\sum_{i=1}^{k} X_i^2}.$$

De la mateixa manera que abans, $\sum_{i=1}^{k} \theta_i X_i$ i $\sum_{i=1}^{k} X_i^2 - k$ tenen la mateixa esperança,

$$\mathbb{E} \sum_{i=1}^{k} \theta_i X_i = \sum_{i=1}^{k} \mathbb{E}(\theta_i X_i) = \sum_{i=1}^{k} \theta_i \mathbb{E}(X_i) = \sum_{i=1}^{k} \theta_i^2,$$

$$\mathbb{E}\left(\sum_{i=1}^{k} X_i^2 - k\right) = \mathbb{E}\sum_{i=1}^{k} X_i^2 - \mathbb{E}(k) = \sum_{i=1}^{k} \mathbb{E}(X_i^2) - k = \sum_{i=1}^{k} (\theta_i^2 + 1) - k =$$
$$= \sum_{i=1}^{k} \theta_i^2 + k - k = \sum_{i=1}^{k} \theta_i^2.$$

Per tant, podem donar l'estimador James-Stein de la forma,

$$\hat{\beta}_i^c = \left(1 - \frac{k}{S^2}\right) X_i,$$

amb $c = k$.

Aquest punt de vista "Galtonià"aporta una visió força clara de la paradoxa i de com arribar a deduir els estimadors James-Stein i Efron-Morris. Mentre que l'estimador ordinari és el derivat per la recta de regressió "equivocada", els estimadors James-Stein i Efron-Morris són els derivats per la recta de regressió "correcta". Aquesta visió també ajuda a veure perquè es dóna aquesta situació per a $k \geq 3$. En els casos en que $k = 1$ o $k = 2$, la recta de regressió de θ sobre X obtinguda pel mètode de mínims quadrats, ha de passar necessàriament pels punts (X_i, θ_i). De fet, ambdues rectes de regressió, $X \sim \theta$ i $\theta \sim X$, passaran pel punt X_i i, en conseqüència, l'estimador ordinari estimarà igualment bé θ tot i fer-ho des de la recta de regressió "equivocada".

Conclusions

En aquest treball s'ha fet una revisió històrica del fenomen de la regressió a la mediocritat de Sir Francis Galton. Pare de la regressió i correlació, va ser capaç de veure l'estudi de l'evolució de les espècies com un procés estadístic en el qual l'estudi de l'aleatorietat jugava un paper igual d'important que el de l'heretabilitat. D'aquesta manera va poder donar justificacions matemàtiques a les teories sobre l'evolució de les espècies, les quals estudiava paral·lelament al seu cosí, Charles Darwin. Els nous mètodes introduïts per Galton van suposar el naixement de l'estadística moderna, donant un fort impuls a l'anàlisi i visualització de dades.

A més, el treball posa de manifest quines són les subtileses que envolten el fenomen i fan que aquest sigui susceptible a males interpretacions. És fàcil que, en estudis temporals on l'objecte d'estudi és ambiciós, es passin per alt aquestes fluctuacions estadístiques provinents dels mètodes utilitzats en l'estudi. Tanmateix, mètodes com l'agrupació de dades o la simplificació de grans quantitats de dades a partir de les mitjanes, afavoreixen que el fenomen quedi amagat, generant així, conclusions errònies.

Finalment, seguint l'exposició d'Stigler (1990 [31]), s'ha mostrat com pot interpretar-se la paradoxa d'Stein com un problema de regressió. Aquesta perspectiva aporta claredat a la paradoxa mitjançant el raonament simple i elegant de la regressió i permet desenvolupar una construcció intuïtiva dels estimadors James-Stein i Efron-Morris. Aquests estimadors sovint són presentats com una idea sorprenent força difícil d'acceptar i contraris al pensament intuïtiu. És per aquest motiu que l'argument que es presenta al treball, paral·lel a la rigorosa demostració, aporta llum al tema i permet seguir d'una manera estructurada i ordenada la construcció d'aquests estimadors admissibles vers l'estimador ordinari.

Annex

En aquest annex s'adjunta tota la informació corresponent a les gràfiques i taules que s'han generat especialment per aquest treball.

1. Codi R per a la realització de la Figura 1 del Capítol 1.2 pàgina 3

```
require(psych)
dades.galton<-galton
View(dades.galton)
sunflowerplot(dades.galton$parent,dades.galton$child, xlab = "Alçada dels pares (en polzades)",
     ylab = "Alçada dels fills (en polzades)", pch = 19, col = "black", seg.col="black")
```

2. Codi R per a la realització de la Figura 3 del Capítol 1.3 pàgina 6

```
setwd("C:/Users/julia/Dropbox/Julia.Arago/Docs/ADIP/16.Dades")
dades<-read.table("peas.txt")
View(dades)
summary(dades)

options(digits = 10)
hist(dades$child)
hist(dades$parent)
reg<-lm(dades$child~dades$parent)
sunflowerplot(dades$parent,dades$child, pch=20,
     ylab = "Diàmetre de les llavors filles (en centèssimes de polsada)",
     xlab = "Diàmetre de les llavors mare (en centèssimes de polsada)", col="darkgrey",
     seg.col = "darkgrey")
abline(reg, lwd=2, col="darkgrey")

m21<-mean(dades[1:100,2])
m20<-mean(dades[101:200,2])
m19<-mean(dades[201:300,2])
m18<-mean(dades[301:400,2])
m17<-mean(dades[401:500,2])
m16<-mean(dades[501:600,2])
m15<-mean(dades[601:700,2])

points(15,m15, pch=20, col="red")
points(16,m16, pch=20, col="red")
points(17,m17, pch=20, col="red")
points(18,m18, pch=20, col="red")
points(19,m19, pch=20, col="red")
points(20,m20, pch=20, col="red")
points(21,m21, pch=20, col="red")

vx<-c(15,16,17,18,19,20,21)
vy<-c(m15,m16,m17,m18,m19,m20,m21)

mean.reg<-lm(vy~vx)
abline(mean.reg, col="blue")

parent.var<-var(dades$parent)
parent.var

child.var<-var(dades$child)
child.var
```

3. Codi R per a la realització de les Figures 7 i 8 del Capítol 3.3 pàgines 23-24

Per tal de construir les Figures 7 i 8, s'han utilitzat les dades de la temperatura mitjana del mes de gener del 2016 i 2017, diverses estacions meteorològiques de Catalunya, mesurades en graus centígrads, extretes del web del *Servei Meteorològic de Catalunya*: `http://www.meteo.cat/wpweb/climatologia/serveis-i-dades-climatiques/anuaris-de-dades-meteorologiques/xarxa-destacions-meteorologiques-automatiques/`

Concretament, per a la realització de la Figura 7 s'han utilitzat les dades de les 12 ciutats de la comarca del Segrià. Es mostren a continuació a la següent taula,

	Gener 2016	Gener 2017
Aitona	7.9	3.8
Alcarràs	6.9	3.5
Alfarràs	6.9	3.3
Alguaire	6.9	3.2
Els Alamús	7	3.3
Gimenells	7.2	3.5
Lleida	7.3	3.8
Maials	7.5	3.9
Raimat	7	3.2
Seròs	8	3.8
Torres de Segre	8	3.9
Vilanova de Segrià	6.9	3.3

Per realitzar la Figura 8, s'han utilitzat les temperatures de 8 ciutats d'arreu de Catalunya agafades de nord a sud, recollides a la següent taula:

	Gener 2016	Gener 2017
Vielha	4.4	-0.3
El Pont de Suert	3.4	0.5
Tremp	5	1.8
Camarasa	5.9	2.6
Lleida	7.3	3.8
Horta de Sant Joan	8.4	5
El Parelló	10.9	7
Amposta	11.1	8.3

El Codi utilitzat ha estat el següent:

```
#Scatterplot per a 12 ciutats de la comarca del Segria
#x1: temperatures mitjanes de Gener de 2016
#y1: temperatures mitjanes de Gener de 2017
ciutat1<-c("Aitona", "Alcarras", "Alfarras", "Alguaire", "Alamus", "Gimenells", "Lleida",
           "Maials", "Raimat", "Seros", "Tores de Segre", "Vilanova de Segria")
x1<-c(7.9, 6.9, 6.9, 6.9, 7.0, 7.2, 7.3, 7.5, 7.0, 8.0, 8.0, 6.9)
y1<-c(3.8, 3.5, 3.3, 3.2, 3.3, 3.5, 3.8, 3.9, 3.2, 3.8, 3.9, 3.3)

dades1<-data.frame(ciutat1, x1, y1)

plot(dades1$x1, dades1$y1, xlab = "Gener 2016", ylab = "Gener 2017", pch=19,
     col="purple", ylim=c(3,4.1), xlim=c(6.5,8.3))
#pos=3 -> a dalt
text(x1[6:8], y1[6:8], labels=ciutat1[6:8], cex= 1, pos=3)
text(x1[11], y1[11], labels = ciutat1[11], pos = 3)
text(x1[2], y1[2], labels = ciutat1[2], pos = 3)

#pos=2 -> esquerra
text(x1[1], y1[1], labels = ciutat1[1], pos = 2)
text(x1[3:4], y1[3:4], labels = ciutat1[3:4], pos = 2)
text(x1[12], y1[12]+0.05, labels = ciutat1[12], pos = 2)

#pos=4 -> dreta
text(x1[5], y1[5], labels = ciutat1[5], pos = 4)
text(x1[9:10], y1[9:10], labels = ciutat1[9:10], pos = 4)

#Scatterplot per a 8 ciutats de tot Catalunya
#x2: temperatures mitjanes de Gener de 2016
#y2: temperatures mitjanes de Gener de 2017
ciutat2<-c("Vielha", "Pont de Suert", "Tremp", "Camarasa", "Lleida", "Horta de Sant Joan", "Perellà", "Amposta")
x2<-c(4.4, 3.4, 5.0, 5.9, 7.3, 8.4, 10.9, 11.1)
y2<-c(-0.3, 0.5, 1.8, 2.6, 3.8, 5.0, 7.0, 8.3)
dades2<-data.frame(ciutat2, x2, y2)
View(dades2)
```

```
plot(dades2$x2, dades2$y2, xlab = "Gener 2016", ylab = "Gener 2017", pch=19,
     col="purple", ylim=c(-0.3,8.5), xlim=c(2.9,11.5))
text(x2, y2, labels=ciutat2, cex= 1, pos=3)
```

4. Codi R per a la realització de les Taules del Capítol 3.5 pàgines 27-30

```
set.seed(2019)

## Generem una Simulacio amb matriu de var i cov = Sigma
Sigma <- matrix(c(10,3,3,2),2,2)
Sigma
require(MASS)
X<-mvrnorm(n = 1000, rep(0, 2), Sigma)
str(X)

plot(X[,1],X[,2],pch=19,col="darkgrey",cex=0.1,asp=1, xlab="X", ylab="Y")

mx<-mean(X[,1])
my<-mean(X[,2])
points(mx,my, pch=19, col="black")

## Calcul del coeficient la recta de regresio de Y sobre X: y=a+bx
reg.YX<-lm(X[,2]~X[,1])
reg.YX
abline(reg.YX,lwd=2,col="darkgrey")

## Calcul dels coeficients de la recta de regresio de Y sobre X: x=c+dy
reg.XY<-lm(X[,1]~X[,2])

coef.reg.XY<-reg.XY$coefficients
c<-coef.reg.XY[1]
d<-coef.reg.XY[2]

a1<-(-c/d)
a1
b1<-1/d
b1
abline(a1,b1,lwd=2,col="DarkRed")

## plot en versió blanc i negre
plot(X[,1],X[,2],pch=19,cex=0.1,asp=1, xlab="X", ylab="Y")
abline(reg.YX,lwd=2)
abline(a1,b1,lwd=2, lty="dashed")

## creo un data frame
dades<-data.frame("X"=X[,1], "Y"=X[,2])

#################################### per les X ########################

#ordeno el data frame en funcio dels valors de les X
dades.sorted.X<-dades[order(dades$X),]
View(dades.sorted.X[1:20,]) #percentil 2
View(dades.sorted.X[981:1000,]) #percentil 98
View(dades.sorted.X[491:510,]) #percentil 49 a 51

#funcio que ens dona per un valor concret de X a quin percentil pertany la Y
percentil.X<-ecdf(dades$Y)
percentil.X(dades.sorted.X[1:20,2])*100 # referent al percentil 2 de la X
percentil.X(dades.sorted.X[981:1000,2])*100 # referent al percentil 98 de la X
percentil.X(dades.sorted.X[491:510,2])*100 # referent al percentil 49 al 51

#################################### per les Y ########################

#ordeno el data frame en funcio dels valors de les Y
dades.sorted.Y<-dades[order(dades$Y),]
View(dades.sorted.Y[1:20,]) #percentil 2
View(dades.sorted.Y[981:1000,]) #percentil 98
View(dades.sorted.Y[491:510,]) #percentil 49 a 51

#funcio que ens dona per un valor concret de Y a quin percentil pertany la X
percentil.Y<-ecdf(dades$X)
percentil.Y(dades.sorted.Y[1:20,1])*100 # referent al percentil 2 de la Y
percentil.Y(dades.sorted.Y[981:1000,1])*100 # referent al percentil 98 de la Y
percentil.Y(dades.sorted.Y[491:510,1])*100 # referent al percentil 49 a 51
```

Referències

[1] Cassella, G., Berger, R. L. (2002), *Statistical Inference*, Second Edition, Thomson Learning, Duxbury.

[2] Cortés, R. (1998), *Simulación de la paradoja de Stein con la hoja de cálculo*, Suma, No. 28, pp. 103-107.

[3] Cuadras, C. M. (1991), *Ejemplos y aplicaciones insólitas en regresión y correlación*, Qüestió, Vol. 15, pp. 367-382.

[4] Cutter, G. R. (1976), *Some Examples for Teaching Regression Toward the Mean from a Sampling Viewpoint*. The American Statistician, Vol. 30, No. 4, pp. 194-197.

[5] Efron, B. (2010), *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing and Prediction*, Chapter 1: Empirical Bayes and the James-Stein Estimator, Stanford University.

[6] Elder, R. F. (1934), *Review of The Triumph of Mediocrity in Business by Horace Secrist*, American Economic Review Vol. 24, No. 3, pp. 121-122.

[7] Ellenberg, J. (2014), *How Not to Be Wrong. The Power of Mathematical Thinking*, Part 4, Penguin USA.

[8] Friedman, M. (1992), *Do Old Fallacies Ever Die?*, Journal of Economic Literature, Vol. 30, No. 4, pp. 2129-2132.

[9] Galton, F. (1886), *Regression towards mediocrity in hereditary stature*, Journal of the Antropological Institute, No. 15, pp. 246-263. `http://galton.org/essays/1880-1889/galton-1886-jaigi-regression-stature.pdf`

[10] Galton, F. (1894), *Natural Inheritance*, Macmillan.

[11] González, J. J. (2009), *Regresión a la Media: Un Fenómeno Estadístico con Historia y Repercusión Social*, Universidad de Las Palmas de Gran Canaria.

[12] Gupta, A. K., Peña, E. A. (1991), *A simple motivation for James-Stein estimators*, Statistics & Probability Letters, No. 12, pp. 337-340.

[13] Halliday, T. M., Thomas, D. M., Siu, C. O., Allision, D. B., (2018), Letter to the editor, Journal of Women & Aging, Vol. 30, No. 1, pp. 2-5.

[14] Hanley, J. A. (2004), *"Transmuting" Women into Men: Galton's Family Data on Human Stature*, The American Statistician, Vol. 58, No. 3, pp.237-243.

[15] Hotelling, H. (1933), review of Secrist, H. *The Triumph of Mediocrity in Business*, Journal of the American Statistical Association, Vol. 28, No. 184, pp. 463-465.

[16] Hotelling, H. (1934), addendum to review of Secrist, H. *The Triumph of Mediocrity in Business by Horace Secrist*, Journal of the American Statistical Association, Vol. 29, No. 186, pp. 196-199.

[17] Kenneth, E. J. (1973), *Regression toward the Mean in Uncontrolled Clinical Studies*, Biometrics, Vol. 29, No. 1, pp.121-130, International Biometric Society.

[18] King, W. I. (1934), *Review of The Triumph of Mediocrity in Business by Horace Secrist*, Journal of Political Economy, Vol. 42, No. 3, pp. 398-400.

[19] Koenker, R. (2016), *The Regression Fallacy*, Economics 536, Lecture 8, Department of Economics, University of Illinois. `http://www.econ.uiuc.edu/~roger/courses/508/lectures/L8.pdf`

[20] Monleón, T. (2010), *Importancia de Darwin en el desarrollo de la estadística moderna*, Estadística Española, Vol. 52, No. 175, pp. 371-391.

[21] Pollock, D. S. G. (2011), *Conditional Expectations and Regression Analysis*, Lecture 1, Econometrics EC3062, University of Leicester.`https://www.le.ac.uk/users/dsgp1/COURSES/THIRDMET/MYLECTURES/1REGRESS.pdf`

[22] Rao, S. R. (1973), *Linear Statistical Inference and its Applications*, Second Edition, Pennsylvania State University.

[23] Riegel, R. (1933), *Review of The Triumph of Mediocrity in Business by Horace Secrist*, Annals of the Academy of Political and Social Science, No. 170, pp. 178-179.

[24] Ruiz, G. G. (2003), *Los orígenes del método de mínimos cuadrados*, Suma, No. 43, pp. 31-37

[25] Secrist, H. (1934), reply to review of Hotelling, H. *The Triumph of Mediocrity in Business*, Journal of the American Statistical Association, Vol. 29, No. 186, pp. 196-199.

[26] Smith, G. (2014), *Standard deviations: Flawed Assumptions, Tortured Data, and Other Ways to Lie with Statistics*, Chapter 9: Regression, Gildan Media LLC.

[27] Stanton, J. M. (2001), *Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors*, Syracuse University, Journal of Statistics Education, Vol. 9, No. 3.

[28] Stein, C., James, W. (1961), *Estimation with quadratic loss*, Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, pp. 361-379, University of California Press, Berkeley and Los Angeles.

[29] Stigler, S. M. (1977), *An Attack on Gauss, Published by Legendre in 1820*, Historia Mathematica No. 4, pp 31-35, Academic Press.

[30] Stigler, S. M. (1981), University of Chicago, *Gauss and the Invention of Least Squares*, The Annals of Statistics, Vol. 9, No. 3, pp. 465-474.

[31] Stigler, S. M. (1990), *The 1988 Neyman Memorial Lecture: A Galtonian Perspective on Shrinkage Estimators*, Statistical Science, Vol. 5, No. 1, pp. 147-155, Institute of Mathematical Statistics.

[32] Stigler, S. M. (1996), *The History of Statistics in 1933*, Statistical Science, Vol. 11, No. 3, pp. 244-252, Institute of Mathematical Statistics.

[33] Stigler, S. M. (1997), *Regression towards the mean, historically considered*, Statistical Methods in Medical Research, Vol. 6, pp. 103-114.

[34] Tomisek, A., Flinn, B., Balsky, T., Gruman, C., Rizer, A. M., (2017) *Strong, healthy, energized: Striving for a healthy weight in an older lesbian population*, Journal of Women & Aging, Vol. 29, No. 3, pp. 230-242.